5-2007

Effects of state tests in Ohio on assessment practices in mathematics education.

Brian Thomas Boyd
University of Louisville

EFFECTS OF STATE TESTS IN OHIO ON ASSESSMENT PRACTICES IN MATHEMATICS EDUCATION

By

Brian Thomas Boyd
B.S., Miami University, 1994
M.S., University of Dayton, 1997

A Dissertation
Submitted to the Faculty of the
Graduate School of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

Department of Teaching and Learning
College of Education and Human Development
University of Louisville
Louisville, Kentucky

May 2007

EFFECTS OF STATE TESTS IN OHIO ON ASSESSMENT PRACTICES IN

MATHEMATICS EDUCATION

By

Brian Boyd
B.S., Miami University, 1994
M.S., University of Dayton, 1997

A Dissertation Approved on

March 1, 2007

by the following Dissertation Committee

_____
Dissertation Director

_____

_____

_____

_____

DEDICATION

This dissertation is dedicated to my family, including my parents, Tom and Chris Boyd, my wife, Angie, and our three wonderful children, Brandon, Allison, and Cameron. Thank you, Angie, for all of your support that you have given me to complete this work and pursue my dreams.

ACKNOWLEDGMENTS

I would like to thank my chair, Dr. William S. Bush for all of his support that he provided throughout the PhD program. Thanks also to the other committee members, Dr. Skip Kifer, Dr. Maggie McGatha, Dr. Bob Ronau, and Dr. Chuck Thompson, for their assistance. I would also like to acknowledge all other ACCLAIM faculty members, who not only impacted my education through this program, but had the vision, determination, courage, and persistence to make ACCLAIM a possibility. Thanks especially to Dr. William S. Bush and Dr. Vena Long for navigating through the many challenges you faced to see ACCLAIM come to fruition. This program has changed my life.

I also want to thank the staff of Brookville Middle School, whom I had the pleasure of working with during most of the PhD program. Thanks especially to Tim Hopkins for supporting me in this endeavor, and to Belinda Linville for her help with the classroom assessment data.

Lastly, I want to acknowledge the 13 other members of our ACCLAIM cohort that began in the summer of 2002: Caroline Best, Barbara Buckner, Frank Edge, Craig Green, Julianna Gregory, Debbie King, Brenda Lackey, Bill Maxon, Sue Nichols, Christie Perry, Judy Price-Pennington, Crystal Rice, and Karla Willis. I have enjoyed working with each of you since our first summer together, and this accomplishment means so much more to me because you were a part of it.

ABSTRACT

EFFECTS OF STATE TESTS IN OHIO ON ASSESSMENT PRACTICES IN MATHEMATICS EDUCATION

Brian Boyd

February 5, 2007

Classroom tests from nine eighth-grade mathematics teachers were collected from the 2003-04 and 2005-06 school years. These years represent one school year prior to the eighth-grade Ohio Achievement Test (OAT) in mathematics being implemented and the year after the eighth-grade OAT in mathematics was implemented, respectively. In addition, teachers were interviewed to gain insight into their classroom assessment practices. Classroom assessment data were compared between the two years, and interview data were examined, to investigate the impact that the new state test was having on classroom assessment practices. Teachers increased the percentage of their items that assessing content below the eighth-grade to nearly one-third of their items, and an average of 86% of teachers' classroom assessment items were at the lowest depth of knowledge level during both years. The assessment analysis and interview analysis indicate that teachers' reliance on curriculum materials for their tests, along with teachers' inability to accurately interpret the eighth-grade indicators, are partly responsible for these two findings. The presence of a state test did not entice teachers to assess more eighth-grade mathematics content or higher depths of knowledge.

TABLE OF CONTENTS

LIST OF TABLES

TABLE PAGE

CHAPTER I

INTRODUCTION

With the passage of the Elementary and Secondary Education Act (ESEA) in 2002, all states have implemented accountability systems to document student achievement. States are required to develop academic content standards in the hopes that these rigorous standards will drive classroom curriculum and instruction. States are also required to give annual assessments to measure the degree to which students meet these standards (Lonergan, 2003; United States Department of Education (USDE), 2005). In Ohio, by March of 2006, achievement tests were given to all students in grades 3-8 in reading and mathematics. Mathematics achievement tests were implemented first in eighth grade in the spring of 2005.

State tests are not new in the state of Ohio. In 1990 the Ohio Department of Education (ODE) began administering a proficiency test based on a set of outcomes developed for ninth-grade students. Passing this test was required for high school graduation. The new round of tests, including mathematics tests for eighth grade, are more challenging achievement tests than the previous proficiency tests. They are aligned with Academic Content Standards (ACS) that were developed by ODE and published in 2002. These tests were designed to assess a wider range of mathematics content, as well as more advanced mathematics.

Standards and state assessments have the potential to provide a common focus to teachers. Some believe that if assessments are designed well, they can inform instruction for classroom teachers (Fremer & Wall, 2003; Popham, 1987; Popham, 2003). Some states have attempted to develop tests to encourage the type of curriculum and instruction promoted by professional organizations such as the National Council of Teachers of Mathematics (NCTM) (Cohen & Ball, 1990; Firestone, Mayrowetz, & Fairman, 1998; Schorr, Firestone, & Monfils, 2003; Wilson, 2003). Others believe that testing has negative effects on teaching and learning (Abrams, Pedulla, & Madaus, 2003; Bracey,1987; Corbett & Wilson, 1991; Firestone, Mayrowetz, & Fairman, 1998). Bracey (1987) asserted that tests cannot measure creativity, critical thinking, and persistence, because techniques for large-scale assessments must be convenient and cost effective. Other negative effects of state tests have been documented. For example, many teachers, in attempts to raise test scores, only focus on content on the test (Corbett & Wilson, 1991; Firestone, Mayrowetz, & Fairman, 1998), and administrators sometimes cheat and alter the numbers to make their scores appear better (Darling-Hammond, 2004). Proponents of the testing movement, however, point to reactions of educators as the culprit of these problems, rather than the tests themselves (McBee, 2002; Popham, 1987; Walton & Taylor, 1997). Now that new achievement tests have been implemented in Ohio, it is important to know how these tests affect classroom practices.

Background of Testing

The tradition of standardized testing can be traced back to the time of Horace Mann (1796-1859). Mann was interested in a "common test" that would give objective information about a student's progress, provide information to compare the effectiveness

of teachers and schools, and eliminate the subjectivity of the teacher from the results. The expansion of this common test led to the development of the New York Regents test in 1865 (Gallagher, 2003).

Tests have been prominent throughout American history, mostly for the purpose of sorting students. Development of these tests coincided with the beginning of universal education, which was designed to socialize immigrant children into American society. The findings of Charles Darwin, along with an influx of new immigrants, compulsory school attendance, and restrictive child labor laws were forces that caused scientists to develop better tests to sort students in crowded classrooms (Gallagher, 2003). In the late 19th and early 20th centuries, Edward Thorndike sought to develop a "systematic identification and segregation of students according to their intellectual ability" (Gallagher, 2003, p. 86). He believed that in order to educate men effectively, one must find the men that are most able to be educated.

Alfred Binet developed an intelligence test to identify learning deficiencies in children. His test was used to identify students who did not benefit greatly from schooling, and therefore did not need to be educated at all. Lewis Terman expanded Binet's Intelligence test and renamed it the Stanford-Binet Test of Intelligence. Again, the purpose of this test was to sort people, this time into tracks determined by career paths. Tests like these became a way to socially engineer society. Test results were used to determine what occupation and place an individual should have in life, and a curriculum to develop each track of students emerged in the schools (Gallagher, 2003).

Throughout the 20th century, tests continued to be used to sort individuals. Around the time of World War I, the Army used its Alpha and Beta tests to examine, sort,

train, and discharge nearly two million men. This mass testing gave legitimacy to the practice of using tests to make decisions about people's lives (Hanson, 1993). Schools soon began mass testing to make decisions about children. Terman transformed the Army Alpha test into the National Intelligence Test for schoolchildren in 1919, and over 400,000 copies were sold within a year. In the late 1920's, the Iowa Test of Educational Development (ITED) was produced and was used voluntarily on a state-wide basis. ITED was the most frequently used achievement test in the United States for over 50 years (Gallagher, 2003). The development of aptitude tests, such as the SAT in the 1920s and the ACT in the 1950s, was driven by the need to identify prospective college-bound students. With these tests, colleges and universities dedicated precious educational resources only to those who scored well enough on the tests.

Prior to the recent ESEA, the federal government already expanded standardized testing. In 1965 with the first passage of ESEA, school districts were required to administer standardized tests and submit the results in order to receive federal funding (Gallagher, 2003). In 1969 the federal government supported the development of the *National Assessment of Educational Progress* (NAEP), which since has become known as the Nation's Report Card. In the past two decades, tremendous growth in state-wide testing programs has emerged. Much of this growth was driven by the 1983 report *A Nation at Risk* (Gallagher, 2003). While there has been some dispute over the legitimacy of the claim that Americans were at risk because of failing public schools, fear of this risk combined with society's belief in objective, quantifiable data propelled our nation into a testing frenzy. By 1989, 47 states had adopted policies that expanded state-wide testing programs (Gallagher, 2003).

Standards-Based Education

A Nation at Risk stimulated the standards-based education movement. In 1989 NCTM led other subject matter organizations in publishing its first set of standards for students in grades K-12. These standards described, and provided a rationale for, the mathematics all students should learn. NCTM published *Principles and Standards of School Mathematics* (PSSM), an updated set of standards in 2000. These two standards documents clarified a vision for school mathematics for the purpose of improving mathematics teaching and learning. Much of the content of these documents challenged the traditional notion of school mathematics. They described not only the important mathematics content to be learned by students but also important mathematical processes such as problem solving, reasoning, and communication with mathematics. These documents emphasized learning mathematics with understanding, not just learning memorized facts and isolated skills. They articulated a belief that activities should be engaging and grow out of problem-solving situations. This represented a shift from seeing problem solving as an afterthought to learning mathematics in order to solve problems. (NCTM, 1989). NCTM (1989, 2000) also advocated for all students to have access to quality mathematics curriculum and instruction. This stance challenged the traditional belief that only certain students were capable of learning mathematics and should be exposed to challenging mathematics.

As a result of documents like *A Nation at Risk* and the NCTM Standards documents, states began to develop their own standards and expectations for student learning of mathematics. The recent re-authorization of ESEA in 2002 required all states to develop academic content standards (USDE, 2005). Ohio began the process of

developing their current standards in 1997, culminating in the publication of the Academic Content Standards (ACS) for Mathematics in 2002 (ODE, 2002). Advisory groups and writing teams developed and monitored the creation of Ohio's ACS. These groups were composed of K-12 teachers, higher education faculty, parents, and business leaders. These groups took great care to develop these standards, as they serve as a basis for curriculum, instruction, and assessment in all Ohio classrooms (ODE, 2002). Ohio's ACS in Mathematics closely align with NCTM's PSSM. It included five content standards similar to those of PSSM, and the process standards from PSSM were referenced throughout the document. In addition, the guiding philosophies of Ohio's Standards draw from PSSM (e.g. high expectations in mathematics for all students) (NCTM, 2000; ODE, 2002).

ESEA of 2002 also required states to test students with respect to the standards that each state developed. While the standardized tests described earlier were developed to sort individuals, these new required state tests had a different purpose. Ohio's Achievement Tests (OAT) were used to measure student progress toward the ACS (ODE, 2002). The accountability system, based on these assessments, was intended to allow schools to identify students who were not succeeding so that schools can provide the necessary support (USDE, 2005).

Ohio began high-stakes testing of students in November of 1990. Ninth-grade tests were given in reading, writing, mathematics, and citizenship. These tests were considered "high-stakes" because students were required to pass them in order to graduate from any public high school in Ohio. The ninth-grade tests were based upon a list of learning outcomes adopted by the state board of education in 1988 (ODE, 1990).

Later in the mid-1990's the fourth- and sixth-grade proficiency tests were developed and aligned to a list of outcomes in each content area, similar to the way the ninth-grade test was aligned to its outcomes (ODE, 2006a; ODE, 2006b). Items from the ninth-grade test assessed a very basic knowledge of mathematics. For example, twice as many items fell under the standard of arithmetic as fell in each of the standards of algebra or geometry (ODE, 1990). Also, the test was written so that students were not allowed to use a calculator on the test, and all of the items on the ninth-grade test were multiple-choice questions.

Current Testing Climate

Policy leading to the wider use of standardized tests in the United States has been driven by beliefs that standardized tests, with objective and quantified data, are the best way to evaluate education (Stiggins, 2004). Recently, about two-thirds of the U.S. public believed that the emphasis on standardized testing is at the right level or should be increased (Rose & Gallup, 2001). The role of the federal and state governments in expanding standardized testing is simply a manifestation of the belief of the citizenry. Our society believes that "the path to school improvement is paved with more and better standardized tests" (Stiggins, 2002, p. 22).

Those who support the testing movement see tests as a way to reform education by creating tests that encourage the type of curriculum, instruction, and assessment practices that professional organizations promote (Popham, 1987; Wilson, 2003). Since testing results are public record, many advocates believe that competition will encourage teachers to be more effective.

Not all educators agree that education can be improved through testing. Some believe that the ESEA is having negative effects on public education. While the initial purpose of a test may be to measure student progress toward a set of standards, published test results can lead to sorting of students, teachers, schools, and districts. Prior to the 1970's, standardized testing was viewed more subtly in the United States than it is today. Tests were used by teachers to provide feedback and confirm student learning. The testing outcomes largely remained with teachers and occasionally were shared with parents (Behuniak, 2003). Today, testing results are made available to the general public and are used to promote improvement and competition. When schools fail to improve student learning, or fall behind other schools in the rankings, public support for schools decreases.

New tests also require substantial funding. Prior to ESEA, "no other country in the world (had) as much achievement testing as the United States nor (allowed) such extensive commercial profiteering" (Paris, Lawton, Turner, & Roth, 1991, p. 13). Testing companies are also often responsible for the test-preparation materials that accompany these tests (Paris, Lawton, Turner, & Roth, 1991). More money will be made by testing companies as states continue to develop new tests and comply with the ESEA. For example, in 2004 Measured Progress received a $118 million contract to write the grades 3-8 achievement tests in mathematics and reading for the state of Massachusetts (Olson, 2004). New Hampshire, Vermont, and Rhode Island also have contracted with Measured Progress to develop a common set of tests to be used in all three states (Olson, 2004).

Much is at stake in this testing movement. Tests are used to make decisions about students' education, as well as the effectiveness of teachers, schools, and districts. It is critical that tests have a positive effect on the education of our youth. Therefore, to understand what, if any, effect state tests are having on teaching practices, we must understand what other variables affect teachers' classroom practices.

Influences on Teacher Practice

While tests and accountability systems have been developed to improve teaching practices, many other factors affect the practices of teachers. The beliefs that teachers hold about mathematics, learning, and teaching play important roles in the classroom practices of teachers (Dossey, 1992; Kagan, 1992; Mewborn, 2002; Pajares, 1992). Teachers' knowledge of mathematics has an impact on teachers' practices in the classroom (Ball, 1988; Cooney, Badger, and Wilson, 1993; Ma, 1999; Shulman, 1986). There are also factors within a school setting that influence teaching practices, such as the building principal (Jaberg, Lubinski, & Aeschleman, 2004), other teachers (Taylor, 2004), and the larger school community (Wilson, 2003).

Beliefs About Mathematics

What does it mean to know and do mathematics? The many answers to this question affect what mathematics is taught, how it is taught, and how it is assessed in the classroom (Mewborn, 2002; Rousseau, 2004). Rousseau (2004) studied the attempt and failure of a mathematics department at an urban high school to change the way they taught pre-algebra. Teachers wanted the course to better reflect the vision for school mathematics articulated in NCTM's Standards documents. Throughout the year of Rousseau's (2004) study, teachers struggled with their own traditional beliefs about pre-

algebra and those articulated in NCTM's Standards documents. Creating tasks that were more open ended and that valued reasoning skills and conceptual understanding was difficult for teachers who believed pre-algebra needed to be rules- and skills-based, include much drill and practice, and be very teacher centered. As a result of these conflicting beliefs, high school teachers in this school made no significant changes in the way they taught pre-algebra. Rousseau (2004) believed their failure to change teaching practices was related to their beliefs about mathematics.

Mewborn's (2002) case study examined the beliefs of one particular teacher, Carrie, as well as the structure of her beliefs. Mewborn found that not only are beliefs about teaching, learning, and mathematics important, but also the way they are structured – that some beliefs are derived from more central beliefs. Carrie's core beliefs included respecting students and their abilities to learn. This core belief was at odds with her traditional beliefs about mathematics, that it was a dull subject with no place for enjoyment or creativity. Throughout the study, Carrie's core beliefs about respecting children allowed her to change her beliefs about mathematics so they were more consistent with her core beliefs. Mewborn concluded that teachers' classroom practices are not only dependent upon the content of their beliefs, but also the way those beliefs are structured.

Beliefs about the nature of mathematics tend to fall into two categories (Dossey, 1992). The first is an external view about mathematics. This belief conceptualizes mathematics as a static discipline, with a pre-determined set of knowledge and skills to be studied. Mathematics is viewed as certain, absolute, and value-free. Teachers with this more traditional view of mathematics tend to structure classroom practices around

telling and demonstrating to students the proper technique for specific mathematical exercises, followed by students practicing these procedures (Brown, Cooney, & Jones, 1990; Rousseau, 2004; Thompson, 1992). There is nothing to conceptualize or make sense of, so discovery and sense-making activities are not seen as appropriate strategies for learning mathematics (Brown, Cooney, & Jones, 1990).

The second view of mathematics is an internal view, where one must construct mathematical knowledge and understanding for oneself (Dossey, 1992). This view tends to be considered more often with a reform view of mathematics education (Dossey, 1992). Teachers with an internal belief about the nature of mathematics tend to see doing mathematics as making mathematics (Dossey, 1992; Thompson, 1992). They believe that mathematics not only includes procedural knowledge as described above, but includes understanding of important mathematical concepts. They believe that mathematical knowledge develops out of conjectures, proofs, and refutations, where uncertainty is inherent (Thompson, 1992). Teachers of these classrooms display much different instructional practices. Students in these classrooms tend to be engaged in "purposeful activities that grow out of problem situations, requiring reasoning and creative thinking, gathering and applying information, discovering, inventing, and communicating ideas, and testing those ideas through critical reflection and argumentation" (Thompson, 1992, p. 128).

Beliefs About Learning

Teachers' beliefs about learning can be tied to other strongly held beliefs about students. Mewborn's (2002) case study of a new teacher's beliefs demonstrated how her beliefs about learning were connected to more central beliefs about students. The study

also illustrated how the structure of a teacher's beliefs influenced teaching practices. Mewborn's subject, Carrie, believed that learning is a process. This belief, along with her more central belief about respecting children, guided her teaching practices. Students were encouraged to use multiple processes to answer questions as opposed to using a single correct way. Carrie was comfortable having her students observe her making mistakes, and she valued using wrong answers to help enhance students' understanding. Carrie often posed problems in class that had more than one correct answer, and she often insisted that students explain their answers.

Joram and Gabriele's (1998) research also found that pre-service teachers' beliefs about learning influenced their practice as they began their teaching careers. Their research noted that many pre-service teachers equated learning with interest and motivation. This belief translated into teaching practices that consisted of lessons that focused mostly on making content interesting to students. Students had to be motivated, quiet, and well managed in order for "learning" to occur.

Rousseau (2004) found that teachers' beliefs about learning impacted their teaching practices. Teachers in her study believed that learning, specifically in mathematics, took hard work and discipline. This belief was reflected in the way they approached students about daily homework. They believed some students did not learn because they were not diligent in practicing the skills to learn mathematics.

Beliefs About Teaching

Teachers' beliefs about teaching have significant influence on teacher practice (Mewborn, 2002; Pajares, 1992; Thompson, 1992). Most pre-service teachers' beliefs about classroom practices are well established prior to formally studying to become a

teacher (Ball, 1988; Pajares, 1992). Pajares (1992) shared the story of a child "playing teacher" at an early age, where the child told her doll students to pay attention to a very important lesson. Teaching is different than other professions. Those studying to be lawyers or doctors have not experienced the courtroom or surgery the way that teachers have experienced the classroom previously as students. In addition, many pre-service teachers have been successful in school and therefore have difficulty seeing a need to change teaching practice from what they experienced. All of these factors make changing teachers' beliefs about teaching difficult (Lerman, 1997; Mewborn, 2002).

Teachers' traditional views of teaching can be very consistent with traditional views of mathematics (Brown, Cooney, & Jones, 1990). The traditional view of mathematics fits nicely with a view of teaching that consists of showing or telling students the right way to do mathematics problems. This belief is also consistent with the role of teacher authority in the classroom. The teacher knows the correct way to think about mathematics and explains that correct way to students. These notions of authority outweigh the quest for sense-making in the classroom (Brown, Cooney, & Jones, 1990). In addition, teachers traditionally depend on external authority about the mathematics content that is taught. Textbooks, state tests, or courses of study serve as the source for what content teachers teach. Often, decisions about what content is important are made by those outside of the classroom environment, not by teachers or students. Traditionally, teachers accept this line of authority (Brown, Cooney, and Jones, 1990).

Changing Beliefs

Changing teachers' beliefs about mathematics, learning, and teaching during adulthood is rare (Pajares, 1992). These beliefs develop over a long period of time and

are embedded in our culture. However, Mewborn (2002) credits this lack of change to not enough attention to the structure of beliefs. She distinguishes between primary beliefs, ones that cannot be justified but are more self-evident, and derived beliefs, ones that are logically derived from other beliefs. Mewborn (2002) found success in changing beliefs when examining the content of teachers' beliefs and the structure of those beliefs. When teachers became aware of these beliefs, and the inconsistencies that existed between them, then they were able to make changes to those beliefs.

Teacher Knowledge of Mathematics

Teachers' knowledge of mathematics has significant impact on their practices in the classroom (Ball, 1988; Cooney, Badger, and Wilson, 1993; Ma, 1999; Shulman,1986). Teaching practices have been shown to vary by how teachers' knowledge of mathematics is organized (Ball, 1988; Ma, 1999). The way in which a teacher's knowledge of mathematics is organized is just as important as *how much* knowledge of mathematics a teacher has. Teachers who have deep conceptual understandings of mathematics teach mathematics very differently than those with more basic, superficial understandings (Ma, 1999). Teachers with deep understandings of mathematics are able to create learning opportunities that develop conceptual understanding in their students. They also structure classroom activities in such a way that key concepts build on one another and are connected to one another (Ma, 1999). Teachers' knowledge of mathematics is also connected to their beliefs about mathematics. If teachers believe that mathematics is a set of isolated topics and skills, they tend not to see the importance of developing a conceptual understanding of mathematics.

School Influences

Factors within school settings also effect teacher practice. Building principals, who allocate resources and conduct evaluations of teachers, have great influences over teachers' classroom practices. Principals are considered the instructional leaders of the school (Hoy & Hoy, 2003). They are essential for establishing visions for the schools and establishing climates for important changes to occur. Principals also play important roles in providing teachers with collaborative time to work together on their practice (Jaberg, Lubinski, & Aeschleman, 2004).

Research has shown that collegial work within a school building can influence teachers' classroom practices. Louis, Kruse, and Marks (1996) found that professional communities within a school influenced teachers' pedagogy. Rousseau's (2004) and Taylor's (2004) research demonstrated how collegial work within a school building influenced mathematics teaching practices within that school. Collegial work among teachers can be responsible for changing what and how mathematics is taught and assessed in the classroom (Rousseau, 2004; Taylor, 2004).

Beyond the influence that teachers and principals have within a school, parents can also influence classroom practices. Parents have the ability to place pressure on boards of education, superintendents, principals, and teachers concerning teaching practices. Wilson (2003) cited examples of parents and community groups challenging mathematics educational reform in their schools.

Cultural Influences

As stated earlier, many pre-service teachers' beliefs about classroom practices are well established prior to formally studying to become a teacher (Ball, 1988; Pajares,

1992). These beliefs are not derived from studying explicitly how to teach, but are formed by observation and participation in schooling throughout their childhood (Pajares, 1992; Stigler & Hiebert, 1999). Stigler and Hiebert (1999) believe that these experiences define teaching as a cultural activity. In their analysis of typical classrooms in the United States, Germany, and Japan, they found that teachers in the United States approached classroom teaching with very similar scripts (Stigler & Hiebert, 1999). Parents, community leaders, board members and even teachers share many of these common cultural experiences. These experiences reflect a school community with clear ideas of what school classroom practices should look like. These cultural experiences and beliefs about teaching can have a significant impact on teaching practices (Stigler & Hiebert, 1999). They also explain why teaching practices can be difficult to change and why new practices are not invented each day. Making changes to highly cultural activities such as teaching must evolve over time (Stigler & Hiebert, 1999).

Political influences

Recently, local, state, and federal policies have been directed at improving teaching practices (Cohen & Ball, 1990; ODE, 2002; Wilson, 2003). In the last 50 years, school and educational policies were enacted in response to the launching of Sputnik, the report *A Nation at Risk*, NAEP results, and international studies such as The Third International Mathematics and Science Study (TIMSS). When our educational system fails, policies by federal, state, and local agencies are often enacted to improve schools and teaching practices. These policies have taken the form of developing curriculum and standards for learning, professional development initiatives, and state accountability systems that include state tests. Many of these, including Academic Content Standards,

professional development initiatives, and state tests, currently exist for mathematics in Ohio.

State Test Influences

State assessments are generally designed to measure student progress against grade-level indicators and benchmarks. They have also been used to hold teachers and schools accountable for student performance and encourage improvements in teaching practices. Therefore, state tests must be valid and reliable instruments so that they can allow for comparisons among teachers, schools, and districts.

Proponents of the testing movement view them as ways to improve educational practices and ensure high expectations of all students (Fremer and Wall, 2003; Popham, 1987, 2003; USDE, 2005; Wilson, 2003). Research has shown that state tests have had an impact on classroom practices, but not always positive (Abrams, Pedulla, & Madaus, 2003; Corbett & Wilson, 1991). State tests have resulted in teachers using strategies that are contradictory to good educational practice. Teachers have tended to focus only on content that is on the state test and use assessment items that mirror the format of the state test (Abrams, Pedulla, & Madaus, 2003; Corbett & Wilson, 1991). Unfortunately, many teachers have felt pressure to adopt direct teaching methods to improve test scores (Grant, 2000). Smith (1991) adds that the abundance of state-mandated tests has led to the deskilling of teachers. According to Smith, the reliance on multiple-choice tests also leads to teaching multiple-choice teaching, and teachers are encouraged to teach one, specific way: reduce tasks to their simpler components, and drill students until they master each component. Tests encourage teachers to abandon their own content and pedagogical knowledge for direct telling and drilling methods of teaching. Over time,

teachers become dependent on these methods and this reduced form of teaching becomes the norm (Smith, 1991).

Assessment

Assessment is the process of gathering information about students' progress (NCTM, 1995; NMSA, 2003). Under this broad definition of assessment is a variety of different tasks, approaches, and purposes. Much has been written about the different forms of assessments – from classroom performance tasks and observations of students to objective state-mandated standardized tests. The literature on assessment distinguishes *assessment of learning* and *assessment for learning* (Stiggins, 2002). The next sections describe and discuss the differences in purpose and function of these types of assessments.

Assessment of Learning

Assessment of learning has been the norm of education assessment in the United States for most of the last century (Shepard, 2000; Stiggins, 2002). Traditionally, assessments are given at the end of a chapter, unit, or other learning experience. They are used to check for student learning. This type of assessment practice tends to use items that focus on very isolated skills and facts. This view and practice is consistent with behaviorist theories of learning, in which skills are learned and assessed in isolation from one another (Shepard, 2000). In mathematics, this practice is also based on the notion that learning mathematics was based on mastering a series of isolated bits of knowledge (Schulman, 1996).

Many assessment practices today continue to use assessments to merely document student learning, rather than give insight into students' cognitive processes (Stiggins,

2004). Ohio tests are developed to document student learning for accountability purposes. Items from a test are given summatively to measure student learning, and each item aligns with one specific indicator so achievement can be reported on a student, teacher, school, or district in that specific area.

Assessment for Learning

Assessment used to impact instruction is called formative assessment (NCTM, 2000). In this type of assessment, teachers use information collected from assessments to inform what and how they teach. In reviewing evidence from formative assessments, teachers plan instruction according the needs and weaknesses of the students. Formative assessment is the means to instruction, providing useful information for teachers in planning their teaching (Shepard, 2000). Assessment happens during learning and is actually used to create and enhance learning (NCTM, 2000). In this type of classroom, assessment and instruction are indistinguishable (Shepard, 2000). Every activity in the classroom becomes an opportunity to gain insight into student progress and use that knowledge to teach more effectively.

In reviewing the existing literature, Black and Wiliam (1998) found that by improving formative assessment, teachers can improve student achievement. In fact, many of the studies found that improved formative assessment significantly helped low achievers, narrowed the achievement gap, and improved achievement overall. A more recent study by Wiliam, Lee, Harrison, and Black (2004) investigated the effects of professional development focused on formative assessment on student achievement. The project involved 25 teachers from six different schools. These mathematics and science teachers learned about formative assessment in a series of half-day and full-day

workshops. Instructors followed up by visiting teachers at their schools. Teachers were encouraged and provided support to put formative assessment into practice in their classrooms. The results showed a positive effect on standardized test scores for students in these classrooms. This finding suggested that time devoted to preparing teachers for effective formative assessment was time well spent. Teachers did not have to choose between good assessment practices in the classroom and good testing results.

The study by Wiliam, Lee, Harrison, and Black (2004) was not about formative assessment; it was about supporting teachers to develop their practice. The majority of teachers not only used formative assessment in mathematics classes, but also used it in other classes outside the scope of this research. The teachers reported that it changed their views of themselves as professionals. The belief that standardized tests serve as a way toward school improvement seems to have blinded educators from focusing on the benefits of improving teachers formative assessment practices (Stiggins, 2002; Stiggins, 2004). In addition, the importance of developing teachers as professionals has been ignored (Wiliam, Lee, Harrison, & Black, 2004).

Classroom Assessment

Good classroom assessment practices have many benefits. They provide teachers with feedback about student progress, help teachers plan further instruction, and promote high levels of student learning. These are just some of the recommendations for good classroom assessment practices suggested by the National Council of Teachers of Mathematics (1995, 2000), the Ohio Department of Education (2002), and the National Middle School Association (2003). Teachers should use fewer right-or-wrong type questions and use more assessment tasks that help them better understand student

thinking (NCTM, 2000). Good assessments involve significant mathematics and require students to think and process (Cooney, Badger, & Wilson, 1993). If students are to know that mathematics is more than just answering right-or-wrong type questions, they must be given tasks that encourage them to think and work, and they must be given the appropriate amount of time to do so (Cooney, Badger, & Wilson, 1993). Good assessments also elicit a range of responses and can be solved in multiple ways (Cooney, Badger, & Wilson, 1993). These types of items give teachers insight into student thinking. In doing so, good assessment items help teachers enhance student learning by giving teachers helpful information to plan for instruction (Black & Wiliam, 1998; NCTM, 2000).

Teachers' assessments also reflect the mathematics content that teachers value (Cooney, Badger, & Wilson, 1993; NCTM 2000). If assessments are tied this closely to beliefs, it also explains why teachers' assessment practices are difficult to change (Shepard, 2000). Analyzing teachers' classroom assessments can give insight into the classroom, demonstrating the content and depth of mathematics being taught.

Summary

Teaching and assessing students are complex processes. Teachers are influenced by many factors in this process, including their own beliefs, knowledge, and school setting. Policies such as state standards and tests are also developed to influence teacher practices. State tests in Ohio, as well as tests in other states, are aimed at improving student achievement (USDE, 2005). However, research has shown that state tests actually deskill teachers and limit their ability to improve student learning (Smith, 1991; Shepard, 2000). We also know that teachers tend to have poor skills at writing good

mathematics assessment items that give insight into student thinking (Black & Wiliam, 1998; Cooney, Badger, & Wilson, 1993). In addition, we know that teachers' assessment practices can be most difficult to change (Shepard, 2000). This difficulty may be due to the fact that assessment reflects teachers' beliefs about teaching, learning, and what it means to do mathematics (Cooney, Badger, & Wilson, 1993; NCTM, 2000; Shepard, 2000). As new tests in Ohio become fully implemented, it is important to know the effects that they have on classroom teachers and their assessment practices.

Research Questions

1. How have eighth-grade teachers' assessments changed over the past two years relative to the mathematics content that they assess?
2. How have eighth-grade teachers' assessments changed over the past two years relative to the depth of mathematics knowledge that they assess?
3. To what factors do teachers attribute any changes that may exist in their classroom assessment practices over the last two years?

CHAPTER II

REVIEW OF LITERATURE

Purpose of State Tests

Statewide achievement tests are commonplace as a result of the ESEA re-authorized in 2002. As an accountability system, federal officials and legislators hoped that the achievement data would help teachers, schools, and districts improve student achievement so that no child is left behind (USDE, 2005). In addition, those interested in reforming education believed that state-mandated tests could act as a lever for reform of curriculum and teaching practices (Schorr, Firestone, & Monfils, 2003; Wilson, 2003). So as new standards were developed in various content areas, and at many different grade levels, some believed that the way to ensure that these standards were met was to test all students according to those standards (Darling-Hammond, 2004; Stiggins, 2004). As a result instruction would become more focused. As standards become more valued by students, teachers, and principals, they will in turn influence curriculum and teaching practices. (Fremer & Wall, 2003; Firestone, Mayrowetz, & Fairman, 1998; Popham, 1987).

Some educators believe that if assessments are designed well enough, they will inform instruction for classroom teachers (Fremer & Wall, 2003; Popham, 1987; Popham, 2003). Popham (1987) advocated that teachers view assessment as more than

an afterthought to instruction. Instead, assessment should be designed to help guide learning experiences in a classroom. The design of quality assessments includes a manageable number of standards and at least some items that are extended-response (Popham, 2003). Some educators in California followed this design as they tried to reform the teaching of mathematics in the late 1980's and early 1990's (Cohen & Ball, 1990; Wilson, 2003). New Jersey also created a state assessment to reform teaching (Schorr, Firestone, & Monfils, 2003). In both cases, the belief was that if a test had merit, then teaching to the test would be appropriate. This attitude was also conveyed in the development of the Maine Educational Assessment (MEA).

> Many educators argue that an ideal test would be such that "teaching to it" would be desirable. Teaching to a multiple-choice test uses instructional time to teach skills that are only used during testing. Because open-response questions like those used in the MEA are more direct measure of what we want students to be able to do in testing (and even non-school) situations, however, teaching students how to deal with such questions should be an integral part of instruction (Maine Department of Education, 1994, in Firestone, Mayrowetz, & Fairman, 1998, p. 96).

Using this view of standardized tests, student performance results can provide teachers with information about each student's strengths and weaknesses. With this information, teachers can develop lessons to make sure each student meets state standards. This information also helps school personnel decide where to focus resources. If a specific area is shown to be weak, then a new teaching approach can be developed or

additional professional development for teachers can be provided in that area (USDE, 2005).

Effects of State Tests

State Tests' Influence on Content

Research has shown that state tests influence teaching practices. Abrams, Pedulla, and Madaus (2003) conducted a nationwide survey of teachers to study the effects of state tests. Teachers reported spending more time on tested curriculum, while de-emphasizing content that was not tested. This pattern was the case more often in states with high-stakes tests than in states with low-stakes tests. High-stakes carry consequences for schools or students, such as not being permitted to graduate without a certain score, or punishing a school because of students' performance on the test. Low-stakes tests do not specify consequences tied to student performance, although the testing results may be published in the local newspapers. Firestone, Mayrowetz, and Fairman (1998) studied the effects of state tests in Maryland and Maine. Maryland had a high-stakes test in place; Maine used a low-stakes test. They found that teachers in Maryland tended to focus more heavily on content that was on the test. In addition, teachers reported eliminating content that tended not to be on the test. This practice happened more in Maryland with a high-stakes test, than it did in Maine with a low-stakes test. Other influences included changing the order of what was taught and stressing one area more than others because of the test. The authors contended that it was difficult to isolate this practice as a direct result of the level of "stakes," but it is clear that the mathematics content was affected in both states.

Taylor, Shepard, Kinner, and Rosenthal (2002) found similar results in surveying 1,000 teachers in Colorado. Their study showed that teachers spent less time on non-tested areas, such as science and social studies. Specific to the area of mathematics, their study showed some increase in the number of mathematics topics that were studied. Teachers reported added attention to probability and geometry. Collectively, these studies (Abrams, Pedulla, & Madaus, 2003; Firestone, Mayrowetz, & Fairman, 1998; Taylor, Shepard, Kinner, & Rosenthal, 2002) found that the presence of a state test created a change in how learning was organized as opposed to increasing learning.

State Tests' Influence on Instruction

Studies have also examined the effects tests have on instructional practices. Grant (2000) examined the effects of the pending state tests in New York. One of the purposes of this test was to create more ambitious instruction. Although the test had not been implemented yet, teachers already felt the pressure to use more direct instruction in their teaching. The presence of the state test encouraged more reductive, rather than creative, forms of teaching.

Other studies have shown that the presence of state tests results in teachers spending more time on test-taking skills. Glassnapp, Poggio, and Miller (1991) studied the effects of state tests in Kansas. They found that teachers reported spending more time on activities such as drills, coaching, and practice on items similar to the test. Abrams, Pedulla, and Madaus (2003) and Taylor, Shepard, Kinner, and Rosenthal (2002) found that teachers used instructional time for test-taking skills as opposed to developing a more thorough understanding of the content. Corbett and Wilson (1991) found that teachers in Pennsylvania approached improving students' test scores in a game-like

manner. Teachers felt pressure to find ways to improve test scores but did not feel they were truly improving the quality of instruction. Teachers in these four studies (Abrams, Pedulla, & Madaus, 2003; Corbett & Wilson, 1991; Glassnapp, Poggio, & Miller, 1991; Taylor, Shepard, Kinner, & Rosenthal, 2002) indicated that state tests forced them to teach in ways that would improve test scores, but were contradictory to good educational practice.

Kuppermintz, Shepard, and Linn (2001) reported that the use of test-preparation curriculum and instruction improved scores, but the strategies did not reflect increased learning in the content area beyond the narrow focus of one test. Urdan and Paris (1994) questioned the validity of test scores and increase in student achievement. They referred to the use of test preparation materials as pollution to achievement results, which therefore make comparisons among schools and districts invalid. In some cases, pressure on schools to have good test scores has led to manipulating student placement and even cheating (Darling-Hammond, 2004). Administrators and teachers find ways to "fix" the results to show some improvement of scores and progress in student achievement. Instead of time spent on developing student learning, time is spent creating an illusion of increased achievement (Darling-Hammond, 2004; Urdan & Paris, 1994).

Smith (1991) also reported that teachers began to gravitate toward one correct way to teach – "reduce a task to simpler components and drill it repeatedly until pupils have mastered it" (p. 11). Multiple-choice testing led to multiple-choice teaching, which lead to teachers being less skilled to teach their students (Smith, 1991). The use of large-scale assessments to drive reform in teaching tended to deskill teachers (Shepard, 2000).

A study of the Vermont Portfolio Assessment program (Koretz, Stecher, Klein, & McCaffrey, 1994) showed relationships among state assessments and teaching practices. Researchers surveyed and interviewed a random sample of 80 teachers. Teachers reported that the assessment program was a worthwhile "burden", and that it had positive influences on their teaching practices. Seventy-five percent of the respondents reported having students spend more time applying mathematics to new situations, and more than two-thirds indicated they focused on having students write more in mathematics. Half the teachers reported having students work in pairs or groups more often, and over 40 percent reported being more positive about mathematics. Some reported that even reluctant teachers had changed their teaching practices and that portfolio use had been expanded beyond the area of mathematics in their buildings. This study revealed a strong relationship between state assessments and teachers' reports of changes in classroom curriculum, instruction, and assessment. A portfolio used as a state assessment is not what most typically consider as state tests, even if they include constructed-response items. It is hard to know if the portfolios themselves affected the positive influence on teaching practices, as no other state system or test type was really compared in this study.

As stated earlier, some state tests were implemented to improve instruction (Grant, 2000; Schorr, Firestone, & Monfils, 2003; Wilson, 2003). Some survey studies revealed this finding when teachers self-reported. These studies found that teachers reported a change in practice; however, observations showed that teachers continued to use traditional teaching methods. For example, New Jersey teachers reported using manipulatives more often in their classes, and teachers attributed this change to the New Jersey state tests (Schorr, Firestone, & Monfils, 2003). After observations of these

teachers were conducted, researchers found that while teachers used manipulatives, they did not use them to foster conceptual understanding. This finding reflects an excellent example of teachers adopting different strategies into their already existing practice without truly changing their approach to teaching.

Cohen and Ball (1990) found similar results in their small case study in which teachers incorporated new "tactics" into their own established practices. Teachers in their study used manipulatives and concrete materials. Instead of using them for developing conceptual understanding, however, these materials were used for memorizing rules and procedures. Firestone, Mayrowetz, and Fairman (1998) also found that teachers' reports of changes in their teaching did not align with observations. Maryland's state testing materials included open-ended activities for teachers to use. Instead of using them to help students better understand the main concepts involved in these tasks, teachers used them to have students only practice mathematical procedures. In addition, while the Vermont Portfolio Assessment cited some changes in teacher practices, researchers based this finding only on teachers' self-reported data (Koretz, Stecher, Klein, and McCaffrey, 1994).

Only one study seemed to reveal significant changes in teaching practice. Taylor, Shepard, Kinner, and Rosenthal (2002) found an increase in the number of teachers who asked students to explain answers to mathematics problems. This increase in reasoning and communication was viewed as positive because it was consistent with NCTM's (2000) vision for mathematics teaching.

These studies implied that state tests have had minimal impact on instruction. This finding disputes arguments on both sides of the testing controversy. If tests are

known to have little effect on instruction, then there is no reason to be concerned that teachers use limited teaching strategies because of a state test. Similarly, advocates who claim that mandatory testing can be the lever for the reform of teaching will find their claims unfounded as well. Furthermore, there is little evidence that tests developed with the purpose of improving teaching have truly had such effects. Large-scale, quantitative research is lacking in this area. More research is needed to help better understand the effects that state tests have on classroom practices.

State Tests' Influence on Assessment

Only two studies examined the effects of state tests on assessment practices. Survey data from McMillan, Myran, and Workman (1999) revealed that teachers reported an influence on their assessment practices as a result of the new state test in Virginia. Teachers reported that they tended to use classroom assessment items that mirrored the format of the state test. Abrams, Pedulla, and Madaus (2003) found that teachers across the United States developed classroom assessments that mirrored the format of the state test. This practice was more common in states with high-stakes tests. A larger percentage of teachers from states with high-stakes tests reported using multiple-choice items on a weekly basis.

Black and Wiliam (1998) and Cooney, Badger, and Wilson (1993) found that few teachers were able to produce good assessment items that provided important information about student learning. Assessment items from the teachers studied tended to encourage rote and superficial learning. The type of feedback that teachers provided on classroom assessments seemed to serve more managerial functions at the expense of learning. State-mandated tests not only deskilled teachers' ability to assess students well (Smith,

1991), but teachers were not very accomplished at writing assessment tasks that challenged students as much as the state tests.

Shepard (2000) argued that teachers' assessment practices are difficult to change. While many teachers believe they teach differently, they have not actually changed their beliefs about teaching or mathematics. If teachers view mathematics as sets of procedures, they will most likely continue constructing the same types of classroom assessments. And if teachers examine the state test through this lens of procedures, the test will have little impact on their view of what it means to know and do mathematics. This also explains why some teachers maintain similar assessment procedures and tasks despite using reform curricula.

Summary of Testing Influence

Learning to teach mathematics differently seems to involve more than just creating a state test on which teachers must focus (Cohen & Ball, 1990). Wiliam, Lee, Harrison, and Black (2004) indicate that supporting teachers to develop their practice, specifically in developing high quality formative assessments, has more impact on student learning than implementing state achievement tests. Expecting teachers to teach or assess mathematics differently than the way they learned is complicated. Policies such as state tests and accountability systems are interpreted through teachers' existing beliefs and knowledge of learning and mathematics (Cohen & Ball, 1990). As stated earlier, classroom practices are strongly related to teachers' beliefs and knowledge about learning and mathematics (Dossey, 1992; Mewborn, 2002; Thompson, 1992).

In interviewing New York teachers about the upcoming, more open-ended test, Grant (2000) found that teachers were positive because the test was created with the

purpose of encouraging better instruction. However, teachers were concerned about having sufficient opportunities to grow professionally and learning to teach differently than they were taught. Teachers seemed willing to change instruction, but they recognized a need for much additional support. This study supported the fact that the test alone can not create reform in mathematics education.

In their study of Maine and Maryland teachers, Firestone, Mayrowetz, and Fairman (1998) found only one teacher who exemplified reform teaching, and this transformation began well before the state test was implemented. The teacher participated in professional development opportunities from higher education institutes and workshops. While many teachers in the study referenced these types of professional development programs, only one teacher was affected to the degree that her teaching was truly reformed. Schorr, Firestone, and Monfils (2003) noted only one district in New Jersey where reformed teaching occurred, and it did so well in advance of a state test. The reform was stimulated by a mathematics coordinator who had a strong mathematics background, networked with professional mathematics educators, and worked with a local university to improve mathematics education over many years. These results showed that many factors are necessary to have an impact on classroom curriculum, instruction, and assessment.

Assessing Mathematics Content

One of the focuses of the NCTM Standards documents is to set more ambitious goals for mathematics learned during the middle school years (NCTM, 1989; NCTM, 2000). In the past, some areas of mathematics have been ignored during the middle school years because of too much focus on arithmetic skills (NCTM, 1989). NCTM

(2000) recommended that middle school students learn significant amounts of algebra and geometry. Ohio's Academic Content Standards (ACS) also recommended that middle school students be knowledgeable in more areas of mathematics than just arithmetic, as grade-level indicators were evenly distributed among the five standards that mirror NCTM's Content Standards.

In examining four different states' tests, Webb (1999) considered the breadth of content to be assessed. Webb (1999) found that only a small proportion of states' objectives were assessed on the corresponding state test. In most cases, state tests measured less than 50% of the objectives under each standard. DeBoer (2004), through his work with Project 2061, also found that many state assessments did not assess the specific content found in states' content standards.

Assessing Depth of Mathematics Knowledge

In addition to considering the content assessed in mathematics, it is also important to consider the depth of mathematics knowledge that is assessed. NCTM (1995) advocated for assessing students' full mathematical power and not simply testing students' knowledge of specific tasks and isolated skills. Mathematical power includes students' ability to "explore, conjecture, and reason logically, as well as the ability to use a variety of mathematical methods effectively to solve nonroutine problems" (NCTM, 1989, p. 5). NCTM (2000) also called for assessment to focus on students' conceptual understanding, not just on recall of facts and algorithms. Assessing various depths of mathematics content requires questions and tasks with a variety of complexity. Research has shown that students engaged in tasks with high levels of cognitive demand are more

likely to achieve the type of mathematics power described by NCTM (Stein & Smith, 1998; Stein & Lane, 1996)

Frameworks have been developed to consider the depth of mathematics being assessed. Porter (2002) considered five depth levels in his analysis of state test items and standards. Stein and Smith (1998), in their work with the QUASAR Project (Quantitative Understanding: Amplifying Student Achievement and Reasoning), describe four levels of cognitive demand in considering mathematics assessment items.

In analyzing state tests, Webb (1999) also created a framework for classifying items by the depth of knowledge they assess. Level one includes items that ask students to perform simple, one-step, straightforward algorithms. Key words for level one questions include "identify," "recall," "recognize," "use," and "measure" (Webb, 1999). Level two is labeled as the skill/conceptual level. These questions require students to make decisions about how to approach the problem, and often include more than one step. Level two questions often ask students to explain, describe, or interpret. Some key words that help to identify level two questions include "classify," "organize," "estimate," "make observations," "collect and display data," and "compare data." Webb (1999) cautions educators to not interpret level two questions as only skills. Some reviews interpret skills narrowly to only include numerical and other basic and common skills. Level two includes more complex and less common skills such as visualization and probability (Webb, 1999). Level three questions are labeled as strategic thinking questions. They require more demanding reasoning and complex cognitive demands. Level three examples include questions that ask students to draw conclusions from observations, make conjectures, or to develop logical arguments for concepts. Questions

with more than one answer can be level one or two. However, items with more than one answer that require students to justify their response would most likely be a level three question (Webb, 1999). Level four is labeled extended thinking. This level requires reasoning, planning, and developing over an extended period of time.

NCTM (1995) called for an increase in cognitive demand on performance tasks on standardized tests. Doing so will provide more information about student thinking and achievement (NCTM, 1995) and better assess student full mathematical power (NCTM, 1989). They recognized the importance of items that assess not only a variety of mathematics content, but assess students at a variety of depth of knowledge levels.

Conflicting Forces and Results

Mathematics Content

NCTM (2000) recommended focusing on more significant mathematics in the middle school years, recognizing the abundance of time traditionally spent during these years on arithmetic. State standards, specifically those in Ohio, reflect this vision of NCTM. Despite the intent of these standards documents, state tests seem to have a negative impact on the breadth of content covered in a classroom. Studies from Abrams, Pedulla, and Madaus (2003) and Firestone, Mayrowetz, and Fairman (1998) found less content was taught in classrooms as a result of the implementation of state tests. Webb's (1999) analysis of state tests revealed that in most cases tests assessed less than fifty percent of the state's objectives under each standard. If teachers narrow their content to mirror state tests, their students will miss significant amounts of content. The only study that found an increase in the breadth of mathematics content was Taylor, Shepard,

Kinner, and Rosenthal (2002), in which teachers began teaching more probability and geometry as a result of the state tests.

Depth of Knowledge

Cooney, Badger, and Wilson (1993) define good assessments in mathematics as those that involve significant mathematics, can be answered in a variety of ways, elicit a range of responses, require students to communicate, and stimulate the best possible student performance. These types of assessments give teachers a more complete picture of student progress. These types of tasks would fall into levels two, three, and four of Webb's (1999) framework. Cooney, Badger, and Wilson (1993) also found that teachers tend to write only lower complexity questions, even though the teachers perceive them to be assessing a deep and thorough understanding. Teachers were not competent at writing assessment questions at these higher complexity levels.

Webb (1999) also found that state test items generally did not meet the depth of knowledge stated in the respective state's objectives. Multiple choice and single answer items did not give insight into student understanding of mathematical concepts (Schulman, 1996). Standardized test items tended to focus on isolated skills or facts instead of broad tasks that require students to use a variety of skills or strategies (Black & Wiliam, 1998). State test items, however, dominated the assessment landscape and were used as models for many teachers (Black & Wiliam, 1998). Black and Wiliam's work, along with Webb (1999), confirms that state tests are poor models for encouraging teachers to assess at deeper levels of mathematics knowledge. Therefore, not only do many teachers seem to be poor at writing assessment tasks that assess a deeper knowledge of mathematics (Cooney, Badger, & Wilson, 1993), but state tests items are

poor models to guide teachers toward improvement in this area (Black & Wiliam, 1998; Schulman, 1996; Webb, 1999).

CHAPTER III

METHODS

With the passage of ESEA in 2002, all states were required to give annual assessments to measure the degree to which students met academic content standards (Lonergan, 2003; United States Department of Education (USDE), 2005). In Ohio, mathematics achievement tests were implemented first in eighth grade in the spring of 2005. By March 2006, achievement tests were given to all students in grades 3-8 in reading and mathematics.

Over the past 20 years, much has been written about the effects of state testing on curriculum and instructional practices. Studies found that teachers narrow the curriculum and teach only the content on the test (Firestone, Mayrowetz, & Fairman, 1998; Taylor, Shepard, Kinner, & Rosenthal, 2002). Other studies found that instructional time was spent more on test-taking skills rather than on learning the content emphasized by the test (Abrams, Pedulla, & Madaus, 2003; Corbett & Wilson, 1991; Glassnapp, Poggio, & Miller, 1991; Taylor, Shepard, Kinner, & Rosenthal, 2002). These studies also found that teachers felt pressure to teach in ways that improved test scores, but that were contradictory to good educational instructional practice.

However, many proponents of state testing believe that state tests can be used as leverage to improve instructional practices. Some states developed tests to encourage the

type of curriculum and instruction promoted by professional organizations such as NCTM (Cohen & Ball, 1990; Firestone, Mayrowetz, & Fairman, 1998; Schorr, Firestone, & Monfils, 2003; Wilson, 2003). Proponents of testing also believe that teachers' adverse reactions, not the test themselves, are largely to blame for the negative influences of state tests (McBee, 2002; Popham, 1987; Walton & Taylor, 1997). While the negative results of testing were not the intentions of the new Ohio tests, to know how tests affect assessment practices in the state of Ohio is an important line of inquiry.

Research Questions

1. How have eighth-grade teachers' assessments changed over the past two years relative to the mathematics content that they assess?
2. How have eighth-grade teachers' assessments changed over the past two years relative to the depth of mathematics knowledge that they assess?
3. What factors do teachers attribute to changes that may exist in their classroom assessment practices over the last two years?

Classroom Assessments

This study investigated the effects that statewide testing in Ohio had on assessment practices of eighth-grade mathematics teachers between the 2003-04 and 2005-06 school years. Classroom assessments were collected from nine eighth-grade mathematics teachers. Assessments reflect the content and processes that teachers value and believe are important for students to learn (Cooney, Badger, & Wilson, 1993; NCTM 2000). Because changing assessment practices can be most difficult for teachers (Cooney, Badger, & Wilson, 1993; Shepard, 2000), collecting classroom assessments can help gain insight into the mathematics that teachers believe important.

Previous studies focusing on the effects of state testing found conflicting reports and overstated results (Cohen & Ball, 1990; Corbett & Wilson, 1991; Firestone, Mayrowetz, & Fairman, 1998; Schorr, Firestone, & Monfils; 2003). While teachers reported changes in teaching practices, their actual practices did not always reflect these changes. When studies examined the intricacies of classroom practices, differences between teachers' perception of change and actual changes in teaching practices were found (Cohen & Ball, 1990; Firestone, Mayrowetz, & Fairman, 1998; Schorr, Firestone, & Monfils, 2003). By collecting and analyzing actual classroom assessment items, some of these validity issues will be addressed.

Subjects

Teachers invited to participate in the study were from Montgomery, Preble, Darke, and Miami counties in Ohio. They were largely chosen for convenience. I lived within two hours of all schools and knew administrators at some schools. This proximity and knowledge increased the likelihood that teachers would participate in this study. The administrators helped with response rate, but did not bias the assessment or interview data submitted by teachers. To be included in the study, teachers must have taught eighth-grade mathematics during the 2005-06 school year and must have taught the same level of eighth-grade mathematics course during the 2003-04 school year. During the 2003-04 school year, the new eighth-grade achievement test had not been developed, and little was known about it. The first sample items of the new test were released by the Ohio Department of Education (ODE) in January 2005.

In April 2006, I contacted middle school or elementary principals at all public schools in Montgomery, Preble, Darke, and Miami counties through a letter that

explained the study and asked them to identify qualifying teachers who might be willing to participate. After two weeks, I sent follow-up emails and made phone calls to the principals. A pool of 25 candidates was established from these contacts. However, 16 teachers were unable to participate, mostly because they had not saved assessments from the 2003-04 school year. The remaining nine teachers were a diverse group with respect to the levels of courses they taught, years of experience, type of certification/licensure, and school demographics. I contacted these nine teachers by email and by phone to confirm their participation.

Although teachers in this study were concentrated in one area of Ohio, teachers from multiple counties and schools within the area were included. School districts within the same county often receive similar professional development from the local county educational service center. Teachers from several districts and counties were included because this study focused on the effects of state tests beyond one school, district, county, or professional development opportunity.

Some of the teachers in the study taught more than one level of mathematics at eighth grade. Although their assessment practices may be similar for these different levels, historically teachers have different expectations for some groups of students (NCTM, 2000). While I did not focus on those differences, I acknowledged that they may exist. I collected assessments from both regular and advanced eighth-grade mathematics classes. However, I did not collect assessments from special education courses. I only included teachers specifically trained to be mathematics teachers. Finally, the subjects represented a variety of school demographics, different levels of experience, and different types of licensure or certification.

Table 1

Teacher Profiles

Name	Years	Licensure	Track	Locale/code
Sam	15	7-12	Regular	4
Wanda	6	1-8	Regular	4
Frank	3	4-9	Regular	4
Evelyn	30	7-12, 1-8	Algebra	8
Nancy	35	1-8	Regular	7
Helen	8	1-8	Regular	7
Henry	17	7-12	Regular	4
Edward	9	4-9	Algebra/Integrated I	4
Linda	13	1-8	Regular	4

Table 1 describes each teacher in the study. Pseudonyms are used to maintain confidentiality of the teachers. Four teachers were relatively new to the teaching profession, three were in the middle of their careers, and two were near the end of their careers. Two teachers held licenses to teach grades 7-12 mathematics, four teachers held licenses to teach grades 1-8 in any content area, one teacher held both 7-12 and 1-8 licenses, and two teachers held grades 4-9 licenses to teach mathematics. The teachers taught one of two tracks of classes: (1) a regular eighth-grade mathematics class that was typically considered pre-algebra, or (2) an advanced class that was the equivalent of a ninth-grade course. Seven teachers submitted assessments from their regular mathematics classes, and two teachers submitted tests from their advanced classes.

Edward's advanced class was an algebra class during 2003-04, and it became an integrated mathematics course during the 2005-06 school year, still serving as the equivalent of a ninth-grade course in his district.

The locale codes listed for each school were taken from the United States Department of Education and Institute for Educational Statistics website for classifying school districts on a rural-to-urban continuum (NCES, 2002). Six of the teachers in this sample taught at a school coded 4 – on the fringe of a mid-size city. These districts are typically considered suburban districts. In this case, they surrounded the mid-size city of Dayton. The other three districts were considered rural. The district coded as an 8 was located within the statistical area of a mid-size city, and those coded as 7 were outside of that statistical area. Appendix A has a more complete description of these codes.

Data Collection

All teachers were sent a packet of information that included an explanation of the study and informed consent forms. Some packets included self-addressed stamped envelopes and postage for returning their classroom assessments and consent forms. Others did not include envelopes and postage because I visited the teacher at school personally and gathered these items. All nine teachers also participated in an interview that lasted approximately 20 or 30 minutes. Seven interviews were conducted by phone and two were conducted face-to-face. The interviews were designed to describe (1) changes in assessment practices over the two-year period, (2) the process teachers used to develop assessments and (3) the purpose that assessments serve in their classroom. By the end of the 2005-06 school year, teachers returned all consent forms and tests from the 2003-04 and 2005-06 school years, and all interviews were completed.

Limitations of the Study

In gathering the data, not all classroom assessments were collected. Only classroom tests were collected. In an ideal situation where instruction and assessment were well integrated, it would be difficult to collect every classroom assessment. I assumed that all teachers used some form of tests within their classrooms. While some teachers may have used projects, quizzes, group work, or journals as assessments, all teachers used classroom tests. Focusing only on classroom tests allowed for consistency across the nine teachers. I also felt that the classroom tests would be most likely kept across two years. In fact, some teachers were not selected because they could not locate all of their classroom tests from the 2003-04 school year.

In addition to focusing only on tests, this study was also limited by using only nine teachers. Although teachers were chosen from a variety of counties and types of school districts, the inclusion of only nine limited the generalizations that can be made about the population of teachers in Ohio. However, these teachers and their test items are described in as much detail as possible so that large-scale studies might eventually be conducted. The interview data was intended to further explain factors that affect testing and assessment decisions, and they describe the nature of any changes that occurred over the two years of test data.

Classroom assessment items were coded after interviews were completed with each teacher. Through these interviews, background knowledge was attained about each teacher. This could have led to some potential bias in how individual teachers' items were coded. However, items were coded with attention to the content and depth of knowledge categories. The descriptions for these categories were analyzed throughout

the coding process so that similar or even identical items were consistently coded into the appropriate categories.

Analysis of Assessments

As mentioned earlier, classroom tests were collected from teachers' eighth-grade classrooms from the 2003-04 school year and the 2005-06 school year. Each item of the tests was categorized in two ways. First, items were categorized according to the mathematics content they assessed. Second, items were categorized according to the cognitive demand, or depth of knowledge, that they assessed.

Mathematics Content

The categories for the mathematics content followed directly from the Ohio's Academic Content Standards (ACS). The categories that I used were drawn from the sub-standards of the five standards of Number, Number Sense, and Operations; Measurement; Geometry and Spatial Sense; Patterns, Functions, and Algebra; Data Analysis and Probability. Appendix B shows the sub-standards that served as the content categories, as well as each eighth-grade indicator within each content category. In order for an item to be coded to a content category, the item had to directly assess one of the eighth-grade indicators within that content category.

While most items were placed in only one content category, some items were placed in more than one category. The ACS indicate that teachers should not teach skills in isolation of one another, but that they should be integrated.

> The indicators are presented as separate statements of knowledge, but the intent was not to promote isolated instruction. The indicators should suggest specific content for lessons that will incorporate multiple indicators and cut across the

> standards and disciplines. Effective instruction will not target individual skills alone, but will integrate those skills (ODE, 2002, p. 199).

Therefore, assessment of those skills should mirror connections among content categories. As a result, some items were coded into more than one content category. The number of items that were placed into more than one category, as well as the specific content categories that each item assessed, were recorded. Because some items were coded to multiple content categories, the percentages in the assessment data tables in Appendix C and D may add to more than 100%.

Many items that were coded did not assess content that was considered to be at the eighth-grade level. In order to be coded to an eighth-grade content category, the item had to directly assess one of the eighth-grade indicators. If an item assessed indicators found below the eighth grade, then it was coded as a Number, Geometry, Measurement, Algebra, or Data item that was below eighth grade. Likewise, if an item assessed indicators found above the eighth grade, then it was coded as a Number, Geometry, Measurement, Algebra, or Data item that was above eighth grade. So in addition to the sub-standards listed in Ohio's ACS, each standard was given two additional sub-standards for coding purposes. One was for items that assessed mathematics content considered by Ohio's ACS to be below eighth grade, and one was for items that assessed content considered by Ohio's ACS to be above eighth grade.

For example, an item that required students to find the equation of a line, given only two points, was coded as an Algebra item above the eighth grade. While modeling a description or problem situation with a linear equation was coded as *Use Algebraic Representation* at the eighth grade, the skill of finding the equation of a line given two

points is specifically identified as a ninth-grade indicator. Another example was an item asking students to multiply two monomials or binomials together. This was coded as above the eighth grade because of how the indicators were worded at the eighth- and ninth-grade levels. Multiplying monomials by binomials using "physical models" is mentioned in eighth grade. None of the items coded ever assessed students understanding of multiplying monomials by binomials in this manner. They always narrowly assessed students' skills at simplifying these expressions, which is specifically mentioned in the ninth-grade indicators. Ohio's ACS seem to encourage teachers to teach and assess concepts such as this conceptually before having students use it only as a skill.

Another set of items presenting coding challenges were those involving computation with fractions. Indicators at fifth- and sixth-grade require students to compute fluently with fractions. If an item involved only simplifying number sentences with fractions to be simplified, then it was coded as a Number item below the eighth grade. Items that required students to model situations or solve problems using ratios or rational numbers were coded as eighth grade items under *Computation and Estimation* in the Number standard because of how the indicators were worded in this category. An example might include computing a tip at a restaurant.

Depth of Knowledge

Each item was also categorized according to its cognitive demand, or depth of knowledge. Included in the testing blueprint for the state test are categories for the type of items (multiple choice, short answer, and extended response) as well as the level of

complexity (low, moderate, and high) (ODE, 2004). The blueprint gives the following characteristics for low, medium, and high complexity questions (ODE, 2004).

Low: Requires recall of facts and definitions or performing routine, specified procedures

Medium: Requires some interpretation or multi-step problem solutions.

High: Requires analysis of or justifying reasoning and solution for more complicated problem situations.

The levels of complexity provide descriptions that place items in specific depth of content categories. However, these categories did not sufficiently categorize items in terms of cognitive demand. They also did not provide sufficient descriptions to help decide the complexity level for the assessment items.

The categories that I used for the analysis of items for this study mirror those developed by Webb (1999). In analyzing state tests Webb (1999) created a framework for classifying items by the depth of knowledge they assess. Webb (1999) defined four depth of knowledge categories that an item can assess.

> ***Level 1 (Recall)*** includes the recall of information such as a fact, definition, term, or a simple procedure, as well as performing a simple algorithm or applying a formula. That is, in mathematics a one-step, well-defined, and straight algorithmic procedure should be included at this lowest level. Other key words that signify a Level 1 include "identify," "recall," "recognize," "use," and "measure." Verbs such as "describe" and "explain" could be classified at different levels depending on what is to be described and explained.

Level 2 (Skill/Concept) includes the engagement of some mental processing beyond a habitual response. A Level 2 assessment item requires students to make some decisions as to how to approach the problem or activity, whereas Level 1 requires students to demonstrate a rote response, perform a well-known algorithm, follow a set procedure (like a recipe), or perform a clearly defined series of steps. Keywords that generally distinguish a Level 2 item include "classify," "organize," "estimate," "make observations," "collect and display data," and "compare data."

Level 3 (Strategic Thinking) requires reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. In most instances, requiring students to explain their thinking is a Level 3. Activities that require students to make conjectures are also at this level. The cognitive demands at Level 3 are complex and abstract. The complexity does not result from the fact that there are multiple answers, a possibility for both Levels 1 and 2, but because the task requires more demanding reasoning. An activity, however, that has more than one possible answer and requires students to justify the response they give would most likely be a Level 3. Other Level 3 activities include drawing conclusions from observations; citing evidence and developing a logical argument for concepts; explaining phenomena in terms of concepts; and using concepts to solve problems.

Level 4 (Extended Thinking) requires complex reasoning, planning, developing, and thinking most likely over an extended period of time. The extended time period is not a distinguishing factor if the required work is only

> repetitive and does not require applying significant conceptual understanding and higher-order thinking. For example, if a student has to take the water temperature from a river each day for a month and then construct a graph, this would be classified as a Level 2. However, if the student is to conduct a river study that requires taking into consideration a number of variables, this would be a Level 4. At Level 4, the cognitive demands of the task should be high and the work should be very complex. Students should be required to make several connections–relate ideas *within* the content area or *among* content areas–and have to select on approach among many alternatives on how the situation should be solved, in order to be at this highest level. Level 4 activities include designing and conducting experiments; making connections between a finding and related concepts and phenomena; combining and synthesizing ideas into new concepts, and critiquing experimental designs (Webb, 1999, p. 22-23)

In considering these categories, level 2 items include a wide range of problems. Skill-based problems require less cognitive demand than problems that seek a students' conceptual understanding. Level 1 items require students to demonstrate a rote response, perform a well-known algorithm, follow a set procedure, or perform a clearly defined series of steps (Webb, 1999). For this study routine problems and traditional skill-type problems were coded as level 1 items, while problems that were more conceptual were coded as level 2 items. Some examples of level 1 items that are more skill-based than conceptual are below. These problems seek to assess students' skills and ability to perform algorithms, not students' conceptual understanding of exponents, solving equations, and volume.

- Evaluate: $(2^3)(3^{-2})$
- What is the value of x when $^x/_3 + 5 = 15$
- Compute the volume for a cylinder with radius 5cm and height 25cm.

Webb defines Level 4 as extended thinking. Level 4 items require "complex reasoning, planning, developing, and thinking most likely over an extended period of time" (Webb, 1999, p. 22). No items from the achievement tests or classroom tests included items that qualify as level 4.

Coding Ohio Achievement Tests

In addition to coding classroom tests, I also coded all released items from Ohio Achievement Test (OAT). This included the half-length practice test released in January 2005, released items from the test that was administered in March 2005, and released items from the test administered in March 2006. While all items of the March 2005 test were released, only 19 items (approximately half) were released from the March 2006 test.

Coding Validity and Reliability

In order to ensure that items were coded correctly to the content and depth categories, an eighth-grade mathematics teacher with extensive knowledge of Ohio's ACS also coded items. In addition to coding the assessment items, the teacher was one of the subjects. Her tests were collected to initially develop the framework for coding the items. She submitted her tests well before she knew she would assist in coding the items, and before the categories for coding items were developed. Therefore, data collection from her was no different than from the other teachers. The interview was conducted

before her involvement in the coding process. Furthermore, she did not code any of her own test items.

She and I coded the same sets of items until we reached an acceptable benchmark of 90% agreement. Once we met this benchmark, we coded items individually. We continued to check our consistency in coding throughout our coding of assessment items. After each of us coded a set of assessments, we met for approximately one hour to discuss items that were not clear or may have contradicted earlier decisions about how we coded items. We discussed these items, made notes about how to code similar items, and typed these notes to use later in coding similar items. On several occasions we reviewed items already coded and made changes to ensure continued reliability. We continued this process in coding the assessments of all nine teachers.

Interview Data

All nine teachers also participated in a 20-30 minute interview. The following questions were asked to describe (1) changes in assessment practices over the two-year period, (2) the process teachers used to develop assessments and (3) the purpose that assessments serve in their classroom. The interviews were recorded, transcribed, and analyzed to describe changes that occurred across the two years of testing. In addition, I collected demographic information on each teacher's school district to help ensure a diverse sample of teachers. The questions that served as a guide for the interviews are listed below.

1. Discuss the process that you use to develop mathematics assessments for your students.

a. What resources do you use (tests from curriculum materials, tests from other teachers, state test items) for items?
    i. What curriculum materials are you using this year?
    ii. Is it different than two years ago?
b. How do you decide what mathematics content to test or not test?
c. How do you decide the difficulty level of test questions or tasks?
d. How do you decide formats for test questions or tasks (i.e multiple choice, open-response, fill-in-the-blank)?

2. What is the purpose of giving tests in your classroom? What role do tests serve?
3. Do you assess students' performance or knowledge in mathematics in other ways? What are they?
4. Approximately what percentage of a student's final grade is determined by their test grades?
5. Have you received training or professional development on assessment or developing tests in the past two years? If so, what did you learn?
6. Do you believe that I will notice differences between your tests from two years ago and this year? If so, what would the changes be?

CHAPTER IV

ANALYSIS

This chapter presents the analysis and summary of teachers' classroom assessment items from the 2003-04 and 2005-06 school years. The chapter begins with an analysis of the eighth-grade indicators and released Ohio Achievement Test (OAT) items. Next, teachers' interview data and assessment data are presented individually to fully examine each teacher's assessment practices during the 2003-04 and 2005-06 school years. Comparisons are made between the two years of data for each teacher. These data are also compared to the eighth-grade indicators and OAT released items. Teachers' assessment items and released OAT items are examined, analyzed, and compared in terms of the mathematics content and the depth of knowledge assessed. Interview data are then examined in the context of the teachers' assessment data. Finally, aggregated teacher data are examined to analyze any changes that occurred for the entire sample and like-groups of teachers in the sample.

Eighth-Grade Indicators and Released OAT Items

The eighth-grade indicators were examined in terms of the mathematics content categories that they represent in Ohio's Academic Content Standards (ACS). These indicators serve as the recommended mathematics content from the Ohio Department of Education (ODE). A listing of these indicators and their corresponding content

categories can be found in Appendix B. The released OAT items were also examined for the mathematics content and depth of knowledge assessed. Tables C1-C6 and D1-D6 report the percentages of indicators and released OAT items that aligned with each content category. Tables C7 and D7 report the percentage of each that aligned with the five standards. Tables C10 and D10 report the percentage of released OAT items at each depth of knowledge level. These tables in Appendix C and D report similar data from individual teachers. Indicator and released OAT item data are reported redundantly in these tables to make comparisons to teachers' assessment data easier.

Mathematics Content

Ohio's ACS includes 51 indicators at the eighth-grade. Although indicators are not equivalent and not assessed with the same number of items, the percentages in Tables C1-C6 and D1-D6 reveal how the indicators were distributed among the content categories. The content category with the most indicators was *Use Algebraic Representations*, with 19.6%. Content category *Use Measurement Techniques and Tools* included 15.7% of the indicators, and *Statistical Methods* included 11.8% of the indicators. Since these indicators were at the eighth grade, obviously no indicators were above or below the eighth-grade level. The indicators were somewhat evenly distributed over the five standards; however, almost one-third of the indicators in the eighth grade were in the Algebra Standard (Appendix B).

The most assessed content category for the released OAT items was *Use Algebraic Representations* with 21.1% of the items, followed by *Use Measurement Techniques and Tools* with 18.4% of the items. These percentages of released items were consistent with the indicators, as content categories *Use Algebraic Representations* and

Use Measurement Techniques and Tools also had the largest percentage of indicators. Despite the fact that these items were released from the eighth-grade achievement test, two items (one in geometry and one in data) assessed content considered below the eighth grade. As with the indicators, the released items were evenly distributed among the five standards. Algebra was the most assessed standard with 26.3 % of the items. These data revealed that the released items from the OAT aligned with the eighth-grade content emphasized in the ACS.

Depth of Knowledge

The OAT released items included a higher percentage of recall and skill-type items than one might expect. ODE used levels of complexity to categorize its achievement test items, which are slightly different than the depth of knowledge categories used in this study. The ODE blueprint for the eighth-grade test identified three levels of complexity. Low complexity items required recall of facts and definitions or performing routine, specified procedures. Medium complexity items required some interpretation or multi-step problem solutions. High complexity items required analysis of or justifying reasoning and solution for more complicated problem situations (ODE, 2004).

Webb's (1999) framework for depth of knowledge, which was used in this analysis, included three levels as well. Level 1 included items that ask students to recall factual information or perform straightforward algorithms. Level 2 was labeled as the skill/conceptual level. These items required students to make decisions about how to approach the problem and often asked students to explain, describe, or interpret as part of the task. Level 3 items were labeled as strategic thinking. They required more

demanding reasoning and complex cognitive demands. Level 3 items required students to draw conclusions from observations, make conjectures, or develop logical arguments.

While the complexity levels and depth of knowledge levels are slightly different, the OAT blueprint reported that only 25-35% of the test items required recall of facts or performing routine procedures (ODE, 2004). My analysis revealed otherwise with over 61% of the released items classified as level 1. Level 2 items accounted for 35.5% of the OAT released items, and level 3 accounted for 2.6%. Clearly the released OAT items had a large percentage of level 1 items.

Content-Depth Intersections

Table 2 reports the released items from the OAT in terms of mathematics content and depth of knowledge simultaneously. This table also illustrates the previous findings that algebra was the most assessed standard, the items were fairly evenly distributed among the five standards, and the majority of items were level 1 items. Table 2 also illustrates that algebra level 1 was the most assessed content-depth cell, with 18.4% of all OAT released items. Since such a large percentage of items were at level 1, we would expect that a majority of items in each of the standards were level 1. Table 2 confirms this finding, with the exception of data items. Just over 21% of all released OAT items assessed data content. Almost half of those items, 10.5% of all released items, assessed data content at level 2, while only 9.2% of the items assessed data content at level 1. Table 2 illustrates that each of the other four standards had smaller proportions of level 2 items than the Data Standard. Clearly, data content was assessed with a larger percentage of items that required conceptual understanding.

Table 2

Released OAT items: Mathematics Content and Depth of Knowledge (N = 76)

	Depth of knowledge			
Standards	Level 1	Level 2	Level 3	Totals
Number	11.8%	6.6%	0.0%	18.4%
Measurement	11.8%	7.9%	0.0%	19.7%
Geometry	13.2%	5.3%	0.0%	18.4%
Algebra	18.4%	5.3%	2.6%	26.3%
Data	9.2%	10.5%	1.3%	21.1%
Totals	61.8%	35.5%	2.6%	100.0%

No items assessed number, measurement, or geometry at level 3. In fact, only 2 of the 76 released OAT items were classified at level 3, one of which assessed both algebra and data content. If an item assessed indicators in more than one category, then it was coded to multiple content categories. This explains why the total percentage for level 3 items, as well as other totals in Table 2, may seem incorrect. Table 2 reports that 2.6% (2 items) of the released OAT items were algebra level 3 and 1.3% (1 item) of the items were data level 3. However, since one of the level 3 items assessed both algebra and data, it should not be counted twice. That is why the total for level 3 items is 2.6% (2 items).

Individual Teacher Data

This section analyzes assessment data from individual teachers. Information for each teacher is provided as follows: First, interview data is presented to provide

background and beliefs of each teacher, as well as insight into his or her assessment practices. To review, the following questions served as a guide for the interviews:

7. Discuss the process that you use to develop mathematics assessments for your students.
    1. What resources do you use (tests from curriculum materials, tests from other teachers, state test items) for items?
        a. What curriculum materials are you using this year?
        b. Is it different than two years ago?
    2. How do you decide what mathematics content to test or not test?
    3. How do you decide the difficulty level of test questions or tasks?
    4. How do you decide formats for test questions or tasks (i.e multiple choice, open-response, fill-in-the-blank)?
8. What is the purpose of giving tests in your classroom? What role do tests serve?
9. Do you assess students' performance or knowledge in mathematics in other ways? What are they?
10. Approximately what percentage of a student's final grade is determined by their test grades?
11. Have you received training or professional development on assessment or developing tests in the past two years? If so, what did you learn?
12. Do you believe that I will notice differences between your tests from two years ago and this year? If so, what would the changes be?

Second, data from the teachers' classroom assessment items are presented. The items are examined in terms of mathematics content and depth of knowledge. Assessment data from the 2003-04 and 2005-06 school years are reported and analyzed with regard to changes across years. Lastly, data from four of the nine teachers' assessment items are reported in terms of mathematics content and depth of knowledge simultaneously. Because of changes in their assessment items from 2003-04 to 2005-06, their assessment practices were analyzed further by considering content and depth simultaneously. Five teachers used a very large percentage of level 1 items; therefore, no further examination of content and depth simultaneously was necessary for these teachers.

These sections on each teacher will refer to data found in tables C1-C10 and D1-D10 in Appendices C and D, respectively. Appendix C includes tables that summarize teachers' assessment data from the 2003-04 school year. Tables C1-C6 report the percentage of each teachers' items that were coded to each of the content categories. Table C7 reports the percentage of items coded to each of the five standards. Table C8 reports the percentage of teachers' items, by standard, that assessed content considered below the eighth grade, and Table C9 reports similar percentages for items assessing content considered above the eighth grade. Table C10 reports the percentage of teachers' items that were coded to each of the depth of knowledge categories. Appendix D is organized similarly and includes tables that summarize teachers' data from the 2005-06 school year. Tables D1-D6 report the percentage of each teacher's items that were coded to each of the content categories. Table D7 reports the percentage of items coded to each of the five standards. Table D8 reports the percentage of teachers' items, by standard,

that assessed content considered below the eighth grade, and Table D9 reports similar percentages for items assessing content considered above the eighth grade. Table D10 reports the percentage of teachers' items that were coded to each of the depth of knowledge categories.

Summary data is also reported in each of the tables in Appendix C and D to allow for comparisons to the entire sample, the Ohio ACS, and the released OAT items. Tables C1-C6 and D1-D6 report the percentage of indicators, the percentage of released OAT items, and the average percentage of teachers' assessment items coded to each content category. Tables C7 and D7 report the percentage of indicators, the percentage of released OAT items, and the average percentage of teachers' assessment items coded to each standard. Table C8 and D8 report the average percentage of teachers items that assessed content below the eighth grade in 2003-04 and 2005-06, respectively, while Tables C8 and D9 report similar percentages for items assessing content above the eighth grade. Tables C10 and D10 report the percentage of released OAT items coded to each depth of knowledge category and the average percentage of teachers' assessment items coded to each depth of knowledge category

Sam

Sam was in his 15th year of teaching during the 2005-06 school year. He held a grades 7-12 license to teach mathematics, and he submitted tests from his regular eighth-grade mathematics class. Sam's classroom tests consisted mostly of items from the textbook company. His school used the *Connected Mathematics Project* (CMP) curriculum during both the 2003-04 and 2005-06 school years, and these materials provided a basis for his test items. Sam discussed the importance of writing in

assessment items. He changed the CMP items to include more writing if he thought the items were too simplistic. Sam's purpose for testing was to gain insight into his students' thinking. In his tests, Sam wanted to determine if students could solve problems in multiple ways. Sam also discussed students' use of the graphing calculator; he wanted students to explain how they used the calculator to solve problems. Sam was more interested in the process that students used to answer the items on his tests than the answers themselves. Tests made up 25% of his students' final grades, and he had received no professional development specific to assessment over the last two years. Sam expected small changes in his items over the two years. He indicated that he had added a unit on Statistics during the 2005-06 school year, which should be reflected in comparisons to his 2003-04 items.

Mathematics Content. Tables C1-C3 report the percentages of Sam's assessment items from 2003-04 coded to each content category. The largest percentage (30.9%) of Sam's items assessed content in *Use Algebraic Representations*, and over 81% of all of Sam's assessment items were algebra items. The remaining items were split evenly among number, geometry, and measurement with 6.2% of his items in each standard. None of Sam's items assessed data content. While the Geometry and Measurement Standards each accounted for only 6.2% of his items, all of these items assessed content below the eighth grade. Sam also often assessed *Algebra Below Eighth Grade* with 13.4% and *Algebra Above Eighth Grade* with 17.5%. So although over 81% of his items were algebra items, Tables C8 and C9 reveal that over one-third of them assessed content below or above the eighth-grade level. In fact, Sam's items only assessed 5 of the 15 content categories at the eighth grade during 2003-04.

Tables D1-D3 report the percentages of Sam's assessment items from 2003-04 coded to each content category. Sam's items assessed 6 of the 15 content categories at the eighth-grade level. The categories assessed most often were *Use Algebraic Representations* with 18.7% of his items, *Algebra Above Eighth Grade* with 18.7% and *Use Patterns, Relations, and Functions* with 14.3%. Algebra accounted for 63.7% of Sam's items in 2005-06, and data accounted for 18.7%. Although 3.3% and 6.6% of his items were in geometry and measurement, respectively, all of these items assessed content below eighth grade. In fact, just over 20% of Sam's items assessed content below eighth grade.

In summary, Sam's items assessed one more eighth-grade content category from 2003-04 to 2005-06. While the Algebra Standard was still the most assessed, it decreased from representing 81.4% of his items in 2003-04 to 63.7% of his items in 2005-06. Nearly one-fifth (18.7%) of his items in 2005-06 assessed data content, compared to 0% in 2003-04. Measurement and geometry were assessed minimally in his class during both years, and done so on content considered below eighth grade.

Depth of Knowledge. Table C10 reports that during 2003-04, 61.9% of Sam's items assessed level 1 knowledge, 35.1% assessed level 2 knowledge, and 3.1% assessed level 3 knowledge. Table D10 reports that in 2005-06, Sam's assessments used a slightly smaller percentage of level 1 items at 58.2% and a slightly higher percentage of level 2 items at 38.5%. Sam used the same percentage of level 3 items in 2005-06 with 3.3%. During both 2003-04 and 2005-06, Sam used the lowest percentage of level 1 items and the highest percentage of level 2 items of all teachers in the sample.

Table 3

Sam's Assessment Data, 2003-04: Mathematics Content and Depth of Knowledge

N = 97

	Depth of Knowledge			
Standards	Level 1	Level 2	Level 3	Totals
Number	3.1%	3.1%	0.0%	6.2%
Measurement	3.1%	3.1%	0.0%	6.2%
Geometry	1.0%	5.2%	0.0%	6.2%
Algebra	54.6%	23.7%	3.1%	81.4%
Data	0.0%	0.0%	0.0%	0.0%
Totals	61.9%	35.1%	3.1%	100.0%

Table 4

Sam's Assessment Data, 2005-06: Mathematics Content and Depth of Knowledge

N = 91

	Depth of Knowledge			
Standards	Level 1	Level 2	Level 3	Totals
Number	3.3%	4.4%	0.0%	7.7%
Measurement	1.1%	2.2%	0.0%	3.3%
Geometry	1.1%	5.5%	0.0%	6.6%
Algebra	46.2%	15.4%	2.2%	63.7%
Data	6.6%	11.0%	1.1%	18.7%
Totals	58.2%	38.5%	3.3%	100.0%

Content-Depth Intersections. The analysis also explored mathematics content and depth of knowledge simultaneously. Table 3 reports the percentage of Sam's items from 2003-04 coded to each content-depth cell. Table 4 reports the percentage of Sam's items from 2005-06 coded to each content-depth cell. During 2003-04 the most assessed content-depth cell was algebra level 1, with 54.6% of his items. The second most assessed content-depth cell was algebra level 2, with 23.7%. During 2005-06 algebra items at levels 1 and 2 continued to be a priority for Sam, with 46.2% in algebra level 1 and 15.4% in algebra level 2. No other teachers had comparable large percentages of level 2 items in any one standard.

Tables 3 and 4 also reveal that Sam used more data items in 2005-06, increasing from 0% in 2003-04 to 18.7% in 2005-06. In addition, he assessed this content more at level 2 (11%) than at level 1 (6.6%). Not only did Sam increase the percentage of items in data, he also assessed this content mostly at level 2. Sam was only one of two participants to have standards assessed with larger percentages of level 2 items than level 1 items. These percentages for data items are consistent with released OAT items, where the percentage of level 2 items (10.5%) was greater than that of level 1 items (9.2%).

Summary of Interview and Assessment Data. Sam's interview confirmed some of the findings in his assessment patterns. He added a unit on Statistics in 2005-06. In 2003-04, he had no test items that assessed data content. In 2005-06, however, 18.7% of Sam's items assessed data. In addition, Sam indicated that his goal for assessment was to challenge students and gain insight into their thinking. This goal seemed related to the differences in percentages of level 1 and 2 items. Sam used the lowest percentage of level 1 items in his assessments than any teacher in either year, with 61.9% in 2003-04

and 58.2% in 2005-06. He was the only teacher who used a smaller percentage of recall, memorization, or skill-based items than the released OAT items (61.8%). In addition, Sam had the highest percentage of level 2 items. In terms of overall cognitive demand, Sam's assessments were the most challenging in this sample of teachers.

Wanda

Wanda was in her sixth year of teaching during 2005-06. She held a grades 1-8 license to teach all subjects, and she submitted tests from her regular eighth-grade mathematics classroom. Wanda used a test generator from the publisher of her curriculum materials to develop her tests. She used the same textbook for both 2003-04 and 2005-06. To determine what mathematics to teach and assess, Wanda often began with the indicators in Ohio's ACS and grouped them into units that she could teach. She discussed using extended-response items when she wanted to make items more difficult for students or when she wanted to assess multiple topics at once. Wanda's purpose for testing was to determine what students had mastered. Other assessments that she used included homework and weekly review sheets, which were also used to determine what students had mastered. Wanda attended one summer training from the Ohio Mathematics Academy Program (OMAP) since 2003-04. This training was intended to familiarize teachers with the ACS in grades 7-10; however, Wanda stated that the training was not very helpful. She predicted that there would be few changes in her test items from 2003-04 to 2005-06, as she and her colleagues had worked prior to 2003-04 to prepare for the OAT. She stated that there were changes before this, but not in the two-year period of this study.

Mathematics Content. Table C7 reports that during 2003-04 Wanda's assessment items from her general mathematics class were evenly distributed among the five standards. Her most frequently assessed standard was algebra with 22.2%, and the least assessed standard was Geometry with 17.6%. In addition, Tables C1-C3 report that Wanda's items assessed 12 of the 15 content categories at the eighth-grade level. The content categories assessed most often were *Use Algebraic Representations* with 19.5% and *Measurement Below Eighth Grade* with 13.8%. Interestingly, Table C8 reveals that nearly 40% of her items assessed mathematics content below the eighth-grade level.

Tables D1-D3 reveal that Wanda's assessments covered 11 of the 15 eighth-grade content categories in 2005-06. The category she assessed most often was *Use Algebraic Representations* with 29.1%. The next two most common categories were content categories below the eighth-grade level, *Number Below Eighth Grade* with 11.7% and *Geometry Below Eighth Grade* with 11.2%. In fact, Table D8 confirms that over one-third of all her items assessed content considered below the eighth-grade level. Table D7 reports that almost 40% of Wanda's items assessed algebra content, while the remaining items were evenly distributed among the other four standards.

In 2005-06, Wanda assessed *Use Algebraic Representations* content most frequently, and the Algebra Standard most often. Interestingly, algebra represented a larger percentage of her assessment items, increasing from 22.2% in 2003-04 to almost 40% in 2005-06. As a result, her items were less evenly distributed over the five standards in 2005-06. However, the percentage of items that assessed content considered below the eighth-grade level decreased slightly from almost 40% in 2003-04 to 36.8% in

2005-06. As in 2003-04, she assessed 12 of the 15 content categories at the eighth-grade level in 2005-06.

Depth of Knowledge. Tables C10 and D10 report Wanda's classroom assessment data in terms of depth of knowledge for 2003-04 and 2005-06, respectively. During 2003-04 level 1 items accounted for 90% of Wanda's assessment items, with 9.6% of the items at level 2 and 0.4% at level 3. The depth of knowledge that Wanda's items assessed changed in 2005-06. The percentage of level 1 items decreased from 90% to 83.9%, and the percentage of level 2 items increased to from 9.6% to 16.1%. Wanda used no level 3 items in her 2005-06 assessments. However, considering her change in level 1 and 2 items, Wanda increased the cognitive demand of her items from 2003-04 to 2005-06. She was one of only three teachers to use more than 15% of her items at level 2 in 2005-06.

Content-Depth Intersections. Tables 5 and 6 report the content and depth of Wanda's items simultaneously and allow for further descriptions of her changes from 2003-04 to 2005-06. As stated earlier, Wanda's items were evenly distributed among the five content standards in 2003-04, but algebra accounted for nearly 40% of her items in 2005-06. Furthermore her items were not as evenly distributed among the five standards in 2005-06. Table 5 shows that number level 1 was the most assessed content-depth cell during 2003-04 with 20.3%. Table 6 shows that algebra level 1 became the most assessed content-depth cell in 2005-06 with 35%. Not only did algebra become more of a focus for Wanda in 2005-06, but most of the change was due to an increased percentage of level 1 items.

Table 5

Wanda's Assessment Data, 2003-04: Mathematics Content and Depth of Knowledge

N = 261

	Depth of Knowledge			
Standards	Level 1	Level 2	Level 3	Totals
Number	20.3%	0.0%	0.0%	20.3%
Measurement	18.4%	0.8%	0.0%	19.2%
Geometry	16.5%	1.1%	0.0%	17.6%
Algebra	19.2%	3.1%	0.0%	22.2%
Data	16.5%	4.6%	0.4%	21.5%
Totals	90.0%	9.6%	0.4%	100.0%

Table 6

Wanda's Assessment Data, 2005-06: Mathematics Content and Depth of Knowledge

N = 223

	Depth of Knowledge			
Standards	Level 1	Level 2	Level 3	Totals
Number	15.2%	2.7%	0.0%	17.9%
Measurement	10.3%	3.1%	0.0%	13.5%
Geometry	12.1%	2.7%	0.0%	14.8%
Algebra	35.0%	4.9%	0.0%	39.9%
Data	11.2%	3.1%	0.0%	14.3%
Totals	83.9%	16.1%	0.0%	100.0%

Wanda continued to have a large percentage of level 1 items (83.9%) in 2005-06; however, the percentage of level 1 items decreased and the percentage of level 2 items increased in 2005-06. Tables 5 and 6 show that this increase in level 2 items was distributed evenly among all of the standards except for Data. Data level 2 represented 4.6% of Wanda's items in 2003-04 and just 3.1% of her items in 2005-06. However, the total percentage of data questions decreased from 21.5% in 2003-04 to 14.3% in 2005-06. While the percentage of level 2 data items decreased, level 2 items still represented a larger proportion of her data items in 2005-06 than in 2003-04.

Summary of Interview and Assessment Data. During the interview, Wanda indicated that she did not feel her tests changed from 2003-04 to 2005-06. However, some noticeable changes were evident in the data. In 2003-04 Wanda's items were evenly distributed among the five standards. In 2005-06, however, about 40% of her items assessed algebra content. In addition, her items changed with respect to cognitive demand. Only 9.6% of her items were at level 2 in 2003-04, and this percentage increased to 16.1% in 2005-06.

During the interview, Wanda also discussed using the eighth-grade indicators to determine the mathematics content of her assessments. Many of her items, however, assessed content below the eighth grade. In 2003-04, 39.8% of her items assessed content below the eighth grade, and in 2005-06 this percentage was 36.8%. While this did not change greatly, this finding seems to contradict Wanda's prediction about her assessments being aligned to the eighth-grade indicators in Ohio's ACS.

Frank

Frank was in his third year of teaching during the 2005-06 school year. He held a grades 4-9 license to teach mathematics, and he submitted tests from his regular eighth-grade mathematics classroom. Frank made few decisions about how his tests were developed; most decisions were made by the curriculum leader in his district. The materials used to write items, the mathematics content that was assessed, the difficulty level of items, and the format of the items were all determined by the curriculum leader and based on the indicators and the OAT. Frank's primary purpose in testing students was to determine what students had learned and what concepts they understood. He also stated that test results helped him decide whether or not to re-teach topics or consider a different method of teaching the content for next year. Frank also used other assessments such as quizzes, observations, and informal interviews. Test grades contributed 20-30% of students' final grades in his class. While Frank had no professional development specific to assessment between 2003-04 and 2005-06, he mentioned Baldridge training. This training helped him see testing and assessment as a data-driven process. Frank stated that his test items would look different from 2003-04 to 2005-06 because the 2005-06 items would look more like standardized test items. The curriculum leader made changes in 2005-06 to reflect what the district's teachers wanted. Frank stated that the curriculum leader adjusted the items to better reflect the OAT. While teachers in Frank's district were apprehensive at first about these assessments, Frank reported that they were beginning to see the benefits of this process.

Mathematics Content. Table C7 reveals that Frank's assessment items were somewhat evenly distributed among the five standards in 2003-04. The least often

assessed standard was geometry with 10.9%, and the most often assessed standard was algebra with 35.6%. Tables C8 and C9 reveal that almost all of Frank's items in 2003-04 assessed content at the eighth-grade level. One percent of his assessment items were above grade level, all of them coming from data items, and 3% of his items were below grade level (2% in *Measurement Below Eighth Grade* and 1% in *Number Below Eighth Grade*). Tables C1-C3 reveal that Frank's assessment items also were distributed evenly among the content categories at eighth grade. Only 1 of the 15 eighth-grade content categories, *Spatial Relationships*, was not assessed with his 2003-04 assessment items. In fact this category was not assessed by any of the teachers' items in 2003-04. Content categories *Use Algebraic Representations* and *Use Measurement Techniques and Tools* were Frank's most frequently assessed categories, with 27.7% and 10.9% respectively.

Tables D1-D3 reveal that Frank's items assessed all of the 15 content categories at the eighth-grade level in 2005-06. Half of Frank's items assessed algebra content, while the remaining items were evenly distributed among the other four standards. *Use Algebraic Representations* was the most common category assessed, with 39%. No other content category included more than 8% of the items. In addition, just 17% of Frank's items assessed content considered below the eighth grade. This was the second smallest percentage of any teacher in the sample during 2005-06.

Use Algebraic Representations was Frank's most frequently assessed content category during both years, increasing from 27.7% in 2003-04 to 39% in 2005-06. Algebra was also the most assessed standard during both years and represented a larger percentage in 2005-06 (50%) than in 2003-04 (35.6%). Compared to other teachers in the sample, Frank assessed the largest number of eighth-grade content categories,

assessing 14 in 2003-04 and all 15 in 2005-06. While Frank's items were distributed among all content categories, algebra (and specifically *Use Algebraic Representations*) was emphasized more in 2005-06. The percentage of items that assessed content considered below eighth-grade level increased across the two years, going from 3% in 2003-04 to 17.1% in 2005-06. While still below the mean percentage of all nine teachers (32.5%), this increase is significant.

Depth of Knowledge. Tables C10 and D10 reveal that the depth of knowledge assessed through Frank's items changed significantly between 2003-04 and 2005-06. Frank included 64.4% level 1 items in the 2003-04 assessments, followed by 20.8% level 2 items, and 14.9% level 3 items. This was the highest percentage of level 3 items used by any of the nine teachers in either 2003-04 or 2005-06. Frank's percentage of level 1 items increased to 82.9% in 2005-06, while level 2 and 3 percentages both dropped to 13.7% and 3.4%, respectively. Frank's changes in assessment items, relative to depth level, were the greatest among the nine teachers. While his 3.4% of level 3 items was the highest percentage among other participants in 2005-06, decreasing from 14.9% in 2003-04 represented a significant reduction in the cognitive demand of his test items.

Content-Depth Intersections. Frank demonstrated significant changes over the two years of data. Considering the content and depth of knowledge simultaneously allows for further analysis of these changes. Table 7 reveals that in 2003-04, Frank's assessments heavily emphasized algebra and number, both at depth of knowledge level 1. As noted earlier, Frank had the highest percentage of level 3 items with 14.9% in 2003-04. Table 7 illustrates that most of these questions assessed algebra and data content. Almost 7% of his items were algebra level 3 items and 5% of his items were data level 3

Table 7

Frank's Assessment Data, 2003-04: Mathematics Content and Depth of Knowledge

N = 101

	Depth of Knowledge			
Standards	Level 1	Level 2	Level 3	Totals
Number	19.8%	3.0%	1.0%	23.8%
Measurement	9.9%	3.0%	2.0%	14.9%
Geometry	5.9%	5.0%	0.0%	10.9%
Algebra	23.8%	5.0%	6.9%	35.6%
Data	5.0%	5.0%	5.0%	14.9%
Totals	64.4%	20.8%	14.9%	100.0%

Table 8

Frank's Assessment Data, 2005-06: Mathematics Content and Depth of Knowledge

N = 146

	Depth of Knowledge			
Standards	Level 1	Level 2	Level 3	Totals
Number	17.8%	0.7%	0.0%	18.5%
Measurement	10.3%	0.7%	0.7%	11.6%
Geometry	8.2%	0.7%	0.0%	8.9%
Algebra	39.7%	8.9%	1.4%	50.0%
Data	6.8%	2.7%	1.4%	11.0%
Totals	82.9%	13.7%	3.4%	100.0%

items. In addition, 5% of Frank's items were at level 2 in each of the Geometry, Algebra, and Data Standards. Table 8 shows that Frank's items from 2005-06 were more heavily concentrated in algebra level 1, with almost 40%. His items revealed less emphasis on level 2 and 3 items in 2005-06. Most items in each standard were level 1 items.

Summary of Interview and Assessment Data. Frank predicted that there would be changes in his tests between 2003-04 and 2005-06. This data confirm this conjecture. In 2003-04 only 3% of Frank's items assessed content below the eighth grade; however, this percentage increased to 17.1% in 2005-06. While 17.1% was low compared to the other participants' average (34.4%), the increase from 3% was significant. This finding is noteworthy because all of his tests were created by the curriculum leader to align with the OAT in 2005-06. Frank stated that the content and tests aligned more closely to the indicators and OAT in 2005-06, yet a larger percentage of items assessed content below the eighth-grade level than in 2003-04. In addition, items became less like the OAT in 2005-06 in how they were distributed among the five standards. In 2005-06, algebra was assessed by 50% of his items. Like others in the sample, Frank assessed algebra most often but placed more emphasis than state indicators and the released OAT items called for.

Frank's items also changed significantly with respect to the depth of knowledge levels. In 2003-04, 64.4% of his items were level 1. This percentage increased to 82.9% in 2005-06, which is much higher than the 61.8% of OAT released items. According to Frank, his items in 2005-06 looked more like standardized test items. The data show this resulted in lowering the cognitive demand of the items. Frank's items looked less like the OAT in terms of cognitive demand in 2005-06, despite this being the focus of how his

items were developed. So, clearly the revisions made by the district curriculum leader moved Frank's assessments away from state indicators and tests.

Evelyn

Evelyn was in her 30th year of teaching during the 2005-06 school year. She held a license to teach grades 7-12 mathematics as well as a license to teach grades 1-8. Evelyn submitted tests from her advanced eighth-grade mathematics class, which was equivalent to a ninth-grade algebra class. Evelyn used standardized tests from the textbook company as a starting point for her tests, and she adjusted them occasionally. In 2003-04, her algebra class used a book by DC Heath, while in 2005-06 they used a textbook published by Glencoe. In 2005-06 she also added released OAT items to her tests. Evelyn depended on her textbook for the mathematics content she assessed, although she sometimes switched the order in which the content was taught and assessed. In terms of difficulty, Evelyn wanted her students to be successful, yet challenged. She sought to include items with varying difficulty levels, and her most challenging items were bonus. As for the format of her items, Evelyn preferred not to use multiple-choice. She stated that these were too easy for students. She used short-answer and extended-response items most often and indicated there was a relationship between these forms of items and their difficulty level. Evelyn reported that the purpose of her tests was to check student learning, as well as inform her teaching. She also indicated that tests reveal whether students have the necessary skills. In addition to tests she used worksheets, quizzes, and practice review sheets to assess students. Evelyn discussed the desire to have student complete portfolios or larger projects, but she had not done so yet. Tests accounted for a little less than 50% of students' final grades. Evelyn had received no

professional development specific to assessment over the last two years. While she attended Ohio Council of Teachers of Mathematics (OCTM) conferences, she did not attend any sessions specific to assessment. Evelyn predicted there would be no changes in her test items, other than a few more multiple-choice items in 2005-06. Evelyn disliked multiple-choice items because she wanted to know how students obtained their answers. She also discussed that mathematics was an "in-depth" subject and did not lend itself to being assessed with multiple-choice items.

Mathematics Content. Table C7 reports that a large percentage of Evelyn's assessment items assessed algebra content in 2003-04. Seventy-three percent of her items were coded to this Standard, and the only other standard with a significant percentage of items was number with 18.6%. Although 3.8% of her items assessed geometry content, all of them assessed mathematics content below the eighth grade. In fact, only 40% of Evelyn's items assessed content considered to be at the eighth-grade level, and her items assessed only 6 of the 15 eighth-grade content categories. Tables C1-C3 reveal that Evelyn most often assessed *Algebra Above Eighth Grade* with 45.8% and *Use Algebraic Representations* with 25.4%. This finding is reasonable because Evelyn submitted assessments from her eighth-grade algebra course.

Tables D1-D3 reveal that Evelyn's items assessed 10 of the 15 eighth-grade content categories in 2005-06. Algebra was emphasized in her assessment items, specifically algebra content above the eighth grade. The most assessed categories were *Algebra Above Eighth Grade* with 42.2% and *Use Algebraic Representations* with 27.5%. Algebra was the most assessed standard with over 75% of the items in this standard. Most of these algebra items assessed content considered above the eighth-

grade level. Table D7 reports that Evelyn had very few measurement and geometry items, 10% of her items were classified as data, and 9% were classified as number. Table D8 reports that 13.2% of Evelyn's items assessed content considered below the eighth-grade level. This percentage was the lowest among any teacher in 2005-06.

Evelyn was one of two teachers in the study who submitted tests from advanced classes. In terms of mathematics content, Evelyn's assessments changed very little from 2003-04 to 2005-06. Approximately 75% of her items assessed content in the Algebra Standard in both years. The most assessed content category was *Algebra Above Eighth Grade* during both years as well, with 45.8% in 2003-04, and 42.2% in 2005-06. The remaining items outside of algebra revealed more significant changes. The percentage of number questions decreased from 18.6% to 8.9%, and her percentage of data questions increased from 1.7% to approximately 10%. Evelyn continued to include a small percentage of items that assessed both measurement and geometry questions during both years. Her items assessed more of the content categories at the eighth grade, increasing from 6 in 2003-04 to 10 in 2005-06.

Depth of Knowledge. Almost 98% of Evelyn's items in 2003-04 were level 1. Just over 2% were level 2 and none of her items were level 3. In 2005-06 she used a smaller percentage of level 1 items with 93.1%, and the percentage of level 2 items increased to 6.6%. In addition, 0.3% of the items were level 3, which represented only one item from all 2005-06 classroom tests. While the percentage of level 1 items decreased, Evelyn still assessed using level 1 items over 93% of the time in 2005-06. This percentage is considerably larger than the 61.8% of released OAT items. This percentage is disturbing given the fact that the class was advanced.

Summary of Interview and Assessment Data. Evelyn's prediction was correct regarding no changes in items from 2003-04 to 2005-06. In both years, approximately 75% of the items assessed algebra content. More than half of those items assessed content considered above the eighth-grade level. This finding is reasonable since she submitted items from an algebra class, the equivalent of a ninth-grade course. The items changed very little with respect to depth of knowledge assessed. Level 1 items were predominant, with 97.9% in 2003-04 and 93.1% in 2005-06. Even though over 40% of the items assessed content considered above the eighth grade each year, very few assessed students beyond the memorization, recall, or skill level. Evelyn reported not wanting to use multiple-choice items because they were too easy for students. However, her items assessed students at the lowest depth of knowledge level. This also seems to contradict Evelyn's comments about mathematics being an "in-depth" subject. Very few of her items assessed mathematics in any depth.

Nancy

Nancy was in her 35th year of teaching in 2005-06. She held a license to teach grades 1-8, and she submitted tests from her regular eighth-grade mathematics class. Nancy used tests supplied by the textbook company as the basis for her classroom tests. In both 2003-04 and 2005-06 she used materials from Prentice Hall. However, she wanted to buy new mathematics books that were more aligned to Ohio's ACS. Nancy discussed trying to use open-ended items with students, referring to the 10th-grade Ohio Graduation Test (OGT) items. Interestingly, however, the example of area and perimeter she gave in the interview assessed mathematics content considered below the eighth grade. Nancy also wanted to challenge students but, at the same time, not overwhelm

them. In particular, she wanted to challenge her top students. She commented that she had never been called “mamby-pamby", meaning that her class was challenging. As for the format of her test items, she tried to model the OAT format with similar percentages of multiple-choice, short-answer, and extended-response items. Nancy’s purpose in testing was to assess what students could do without assistance. Cheating was a major concern of hers, and she discussed that tests should accurately represent what each student could do individually. The purpose of her classroom tests was to prepare students mentally for the OAT. As for other forms of assessment, Nancy spoke about using problem solving in class work, quizzes, and homework. She discussed wanting students to think during these assessments and using them to gain insight into their understanding. Nancy identified tests as only one of many assessment tools in her classes, and in fact, tests comprised about 10% of students’ final grades. Nancy attended professional development organized by her district on short-cycle assessments between 2003-04 and 2005-06. Her district had developed a series of tests to administer to students throughout the year that checked their progress toward performing on the OAT. She stated that this experience helped her understand more clearly what the state expected her to teach. The only change that she reported in her tests from 2003-04 to 2005-06 was adding three short-cycle assessments. According to her, other test items were identical for both years of tests.

Mathematics Content. Table C7 reports that Nancy’s 2003-04 assessment items emphasized three of the five standards, number with 38%, algebra with 33%, and data with 19.3%. Measurement and geometry had small percentages of items with 3% and 6.7%, respectively. Table’s C1-C3 reveal that Nancy’s most assessed content categories

were *Number Below Eighth Grade* with 26.7% and *Use Algebraic Representations* with 26%. Over 42% of her items assessed content considered below the eighth grade, and her items assessed 10 of the 15 eighth-grade content categories.

Tables D1-D3 reveal that Nancy's items assessed 13 of the 15 eighth-grade content categories in 2005-06. *Use Algebraic Representations* and *Number Below Eighth Grade* were assessed most often, with 25.1% and 24%, respectively. Over one-third of her items (38.4%) assessed content below the eighth-grade level. Table D7 reports that the Number, Algebra, and Data Standards had significant percentages of items, with 38% in number, 31.5% in algebra, and 18.8% in data. Measurement and geometry were assessed with a small percentage of items in 2005-06.

Nancy used the same tests with her students during both 2003-04 and 2005-06. The only change in testing was that she added a set of three short-cycle assessments in 2005-06, which mostly included released items from the OAT and other items created similar to OAT items. As a result, she assessed more eighth-grade content categories in 2005-06, increasing from 10 in 2003-04 to 13 in 2005-06. This change also resulted in a smaller percentage of assessment items assessing content considered below eighth grade. In 2003-04, 42.7% of her items assessed content below eighth grade, compared to 38.4% in 2005-06. *Use Algebraic Representations* and *Number Below Eighth Grade* were Nancy's most assessed content categories in both years, with approximately half of her items coming from these two categories. She also assessed algebra, number, and data with about 90% of her items during both 2003-04 and 2005-06, while measurement and geometry continued to account for a small percentage of her items.

Depth of Knowledge. Table C10 reports the depth of knowledge levels for Nancy's assessment items for 2003-04. Ninety three percent of her items were at level 1, 7% were at level 2, and no items were at level 3. Since Nancy submitted the same tests in 2005-06, except for the additional short-cycle assessments, Nancy's items showed little change in terms of depth of knowledge. Table D10 reports her percentage of level 1 items decreased to 91%, while levels 2 and 3 increased to 8.7% and 0.3% (only one item), respectively. This large percentage of level 1 items was much higher than the percentage of level 1 released OAT items (61.8%).

Summary of Interview and Assessment Data. As stated earlier, Nancy's test items data changed very little from 2003-04 to 2005-06 with respect to both content and depth of knowledge. Despite reporting that she had a better understanding of what content the state expected her to teach, 38.4% of her items in 2005-06 continued to assess content considered below the eighth grade. This percentage was slightly less than the 42.7% in 2003-04. Her concern that the content of her textbook was not aligned with Ohio's ACS was well founded. However, she did not change the textbook tests over the two years to address this concern, other than adding a few short-cycle assessments to her tests. Nancy's comment about not being a "mamby-pamby" was also interesting, considering that 93% of the test items were level 1 in 2003-04 and 91% of the test items were level 1 in 2005-06. Despite feeling like she challenged students, over 90% of her items were at a very low cognitive demand both years. While she discussed assessing students' thinking and conceptual understanding in other ways in her classrooms, her test items did not reflect this practice.

Helen

Helen was in her eighth year of teaching during the 2005-06 school year. She held a license to teach all subjects in grades 1-8. Helen admitted that mathematics was not one of her strengths. She submitted tests from her regular eighth-grade mathematics class. Helen used tests provided by her curriculum materials, *Transition Mathematics*, for both 2003-04 and 2005-06 school years. During both years she used the same textbook series. Helen indicated that when it came to deciding what mathematics to test, the difficulty level of the items, and the format of the items, she relied on the textbook series. Helen had attempted writing her own items in the past, but students often needed too much clarification with each item. Therefore, she decided to use only tests from the textbook series. Tests in Helen's class served primarily as a way to inform her whether students comprehended what was taught and to review previous content. Helen also used homework, notebooks for organizing worksheets, and vocabulary to assess students. Tests accounted for about one-third of a student's final grade in Helen's class. She received no training between 2003-04 and 2005-06 with respect to assessment, and she reported there were no changes in her tests between 2003-04 and 2005-06.

Mathematics Content. Table C7 report that Helen's 2003-04 assessment items were evenly distributed among the Number, Geometry, and Measurement Standards. Number was assessed with 34.4% of the items while geometry and measurement were assessed with 24.2% and 20% of the items, respectively. Of the nine teachers in the sample, Helen assessed algebra the least often with only 12.1%. Tables C4-C6 reveal that Helen's items assessed 12 of the 15 eighth-grade content areas in 2003-04. However, almost 50% of Helen's items assessed content considered below eighth grade.

Her most assessed content category was *Number Below Eighth Grade* with 27%. Other content areas assessed most often were *Use Algebraic Representations* with 9.8% and *Use Measurement Techniques and Tools* with 9.3%.

Tables D4-D6 reveal that Helen's items in 2005-06 assessed 13 of the 15 content categories at the eighth-grade level. Two of the three most assessed content categories were below the eighth-grade level, with 36% in *Number Below Eighth Grade*, 12.2% in *Use Algebraic Representations*, and 12% in *Geometry Below Eighth Grade*. In fact, nearly two-thirds of the items assessed content considered below the eighth grade. The Number Standard was assessed most often with 42.7%. The remaining items were distributed evenly among the other four standards.

Helen assessed number content most often during both 2003-04 and 2005-06. This Standard accounted for a larger percentage of her items in 2005-06, increasing from 34.4% to 42.7%. While this percentage increased and was larger than the percentage of released OAT items or indicators in number, the other remaining items were more evenly distributed among the other four standards in 2005-06. In addition, Nancy's items assessed 12 of the 15 eighth-grade content categories in 2003-04, and 13 of the 15 in 2005-06. However, a smaller percentage of Helen's items assessed content at the eighth-grade level. In 2003-04, 48.8% of her items assessed content below the eighth grade. In 2005-06, that percentage rose to 63%. *Number Below Eighth Grade* alone increased from 27% in 2003-04 to 36% in 2005-06, meaning that over one-third of her items in 2005-06 assessed number content considered below the eighth-grade level.

Depth of Knowledge. Tables C10 and D10 reveal that the largest percentage of Helen's items was assessed at level 1 in both the 2003-04 and 2005-06 school years.

Helen included 91.2% of her items as level 1 in 2003-04 and the remaining 8.8% were level 2 items, with no items being level 3. Helen's percentage of level 1 items increased to 97.7% in 2005-06, as level 2 items dropped to just 2.1%. Helen had one item, or 0.1%, at level 3 in 2005-06. These large percentages of level 1 items clearly are much larger than the 61.8% of level 1 items found in the released OAT items.

Summary of Interview and Assessment Data. Helen indicated that her tests did not change from 2003-04 to 2005-06; however, her items revealed some change. In terms of content, the percentage of items that assessed content below the eighth grade increased from 48.8% in 2003-04 to 63% in 2005-06. In fact, in 2005-06, over one-third of Helen's items assessed number content below the eighth grade. Helen's percentage of level 1 items also increased. In 2003-04, 91.2% over her items were at level 1. In 2005-06, that percentage rose to 97.7%. While she used the word "concept" in her interview, the depth of knowledge analysis on her tests implied that she valued assessing students' skills and procedures in mathematics. These changes in her test data came mostly from added vocabulary sections on each of her tests. While other parts of her tests were identical from 2003-04 and 2005-06, Helen added a significant number of vocabulary items with each test.

Henry

Henry was in his 17th year of teaching during 2005-06. He held a license to teach grades 7-12 mathematics, and he submitted tests from his regular eighth-grade mathematics class. Henry's mathematics curriculum materials changed between 2003-04 and 2005-06. In 2003-04 he used DC Heath materials, and in 2005-06 he used Glencoe materials. While he used mostly Glencoe tests during 2005-06, he stated that much of

what he taught and tested was driven by the OAT. Henry described deciding what mathematics content to test as "a balancing act." He felt pressure to cover content listed for the OAT and to cover content that the high school teachers wanted. In addition, he felt pressure to cover this material by March, prior to the administration of the OAT. Henry described his tests in terms of difficulty of items as having memorization-type items early in the test, with application-type items in the middle of the test. If he used a challenging task, he often made it a bonus question. Henry chose to use multiple-choice and fill-in-the-blank items most of the time. They were easier to grade and quicker for students to take than open-response items. Henry's purpose in testing students was two-fold: to determine (1) if his students listened and (2) whether they grasped the information. He stated that tests measured how well students knew the material. He also stated that tests were a way to compare students to one another. Henry considered himself a "questioner" in class, and often used questions to assess students. He wanted students to think as opposed to just listening to him talking and answering questions. Through classroom questions, he assessed students by their facial expressions and non-verbal cues. These were the only other types of assessments that Henry discussed in the interview. Henry estimated that 85% of a students grade was determined by tests. The only professional development that Henry received with respect to assessment was focused on the OAT and how to use its data. The only change that Henry mentioned in his assessments between 2003-04 and 2005-06 was that he had fewer tests. Henry mentioned he could not take as much time in 2005-06 to test students, so there should be fewer test items in 2005-06.

Mathematics Content. Table C7 reveals that almost 80% of Henry's 2003-04 items assessed the Algebra and Number Standards, with 47.8% in algebra and 37.1% in number. Tables C4-C6 reveal that his most assessed content categories were *Use Algebraic Representations* with 35.2% and *Number Below Eighth Grade* with 25.7%. Over 40% of Henry's items assessed content considered below eighth grade in 2003-04. In addition, about one-fifth of his algebra items (9.8% of 47.8%) assessed content considered above the eighth grade. Of the 15 eighth-grade content categories, Henry's items assessed 9 of them in 2003-04.

Tables D4-D6 reveal that Henry's 2005-06 items assessed 9 of the 15 content categories at the eighth grade. His most assessed content categories were *Algebra Above Eighth Grade* with 22.4%, *Number Below Eighth Grade* with 21.9%, and *Use Algebraic Representations* with 17.3%. Table D7 reports that the Algebra Standard was assessed most often with 42.1%. Only one-third of Henry's items assessed content at the eighth-grade level. Over 27% assessed content considered above the eighth-grade level, and almost 40% assessed content considered below the eighth-grade level.

Henry's assessments changed very little between 2003-04 and 2005-06 in the content assessed. He continued to assess 9 of the 15 eighth-grade content categories, and approximately 40% of his items continued to assess content below the eighth grade. However, a smaller percentage of his items assessed content at the recommended eighth-grade level, as he assessed more content above the eighth grade. In fact *Algebra Above Eighth Grade* was his most assessed content category in 2005-06 with 22.4%. Algebra was still his most assessed standard, accounting for 47.4% of his items in 2003-04 and 42.1% in 2005-06.

Depth of Knowledge. Tables C10 and D10 reveal that Henry's items showed very little change over two years with respect to depth of knowledge. In 2003-04, 98.3% of Henry's items were level 1. This increased slightly to 98.8% in 2005-06. Level 2 decreased from 1.7% in 2003-04 to 1.2% in 2005-06. Henry used no level 3 items in either year. These percentages of level 1 items were the highest among the nine teachers in both 2003-04 and 2005-06. As a result of using almost all level 1 items in both years, his assessments reflected very little change between 2003-04 and 2005-06.

Summary of Interview and Assessment Data. Henry's prediction regarding administering fewer tests in 2005-06 was correct. In 2003-04, Henry administered 1044 test items. In 2005-06, Henry administered 897 test items. The decrease in number of items was over 10%; however, Henry still had the highest number of items of the nine teachers in the study. All teachers but Henry and Helen submitted between 100-300 items for each school year. With respect to content and depth, Henry's assessments changed very little from 2003-04 to 2005-06. Algebra was the most assessed standard during both years, and over 98% of his items were level 1 during both years. While Henry discussed the OAT as the driver of his tests, Henry's items looked very different than the OAT items in terms of content and depth. Only one-third of Henry's items in 2005-06 assessed eighth-grade content. Almost 40% of his items assessed content below the eighth grade and 27.6% assessed content above the eighth grade. In addition, almost all of Henry's items in both years were level 1, compared to just 61.8% of the released OAT items. Henry did not seem to recognize in the interview how low the cognitive demand of his items was. Henry stated that students today just have too many things on their mind to remember everything. His tests clearly reflected that memory was

important in mathematics, as almost all of his test items were memorization, recall of facts, or skills-based tasks. While Henry characterized himself as a questioner in class, his items were focused more on students' recall, with these items accounting for 85% of students' final grades in Henry's class.

Edward

Edward was in his ninth year of teaching during the 2005-06 school year. He held a grades 4-9 license to teach mathematics and submitted tests from his advanced eighth-grade mathematics class. Like Evelyn's advanced class, this class was equivalent to a ninth-grade mathematics class. Students received high school credit for the course. Edward's textbook and course significantly changed between 2003-04 and 2005-06. In 2003-04 he used McGraw-Hill materials for an algebra course. In 2005-06, he used Glencoe materials for an integrated mathematics course. During 2005-06 Edward developed most of his own test items. Because the materials did not include pre-made tests, he used items from exercises in the book and adjusted them to be like the OAT items. This strategy included items requiring more written responses and explanations from students. His textbook focused heavily on algebra and geometry, so he supplemented the materials to address the other content standards. He also struggled with the mathematics content to teach and test because the course was equivalent to a ninth-grade course. He felt pressure to address eighth-grade content because of the OAT, but he also had to teach ninth-grade content. Edward mentioned that he tested only content that he taught in class. He mentioned some of his college mathematics professors had taught specific mathematics content but then tested different content. Edward wanted to carefully align the mathematics content that he taught and tested. In terms of difficulty,

Edward stated that his tests were easier than his homework assignments. He challenged students more with homework problems so that when they took a test, the content seemed easier. In terms of test format, Edward used no multiple-choice items. His tests included either short-answer or extended-response items. Edward's primary purpose for testing students was to measure how well students grasped the information listed by Ohio's ACS. Tests also provided him with a source of data and feedback to determine if he needed to re-teach topics. Other assessments used by Edward included homework, writing prompts, and manipulatives. Tests accounted for 40-50% of students' grades in his class. While Edward had no professional development specifically in assessment, he participated in Baldridge training. This training aligned with his purpose of testing as a source of data for monitoring student learning. This training also helped him make decisions based on test data, specifically with regard to what mathematics content to teach. Edward stated that his assessment changed from 2003-04 to 2005-06. He and other teachers in his district were much more prepared for the OAT in 2005-06 because they had changed the mathematics content of their tests.

Mathematics Content. Table C7 reveals that over 99% of Edward's 2003-04 items focused on two standards, with almost 86% assessing the Algebra Standard and 13.1% assessing the Number Standard. Edward's assessment items included no geometry items and only a half a percent of his items assessed the Measurement and Data Standards. Tables C4-C6 reveal that Edward's most assessed content categories were *Use Algebraic Representations* with almost 50% and *Algebra Above Eighth Grad*e with 31.5%. The focus on algebra, and specifically on algebra above the eighth-grade level, is reasonable because Edward submitted tests from his algebra class. Sixteen percent of

Edward's items in 2003-04 assessed content considered below the eighth-grade level. This percentage was low compared to the average of 28.5% across all nine teachers. However, because the items concentrated so heavily in algebra and geometry, he assessed only 5 of the 15 content categories at eighth grade.

Tables D4-D6 reveal that Edward's 2005-06 items assessed 10 of the 15 content categories at the eighth-grade level. His most assessed categories were *Use Algebraic Representations* with 24.3%, *Geometry Below Eighth Grade* with 14.1%, and *Number Below Eighth Grade* with 13%. In 2005-06 12.5% of Edwards's items assessed content considered above eighth grade, while 36.5% of his items assessed content considered below eighth grade. The Data Standard was assessed with only 5% of his items, and the remaining items were evenly distributed among the other four standards.

Edward's assessments changed considerably from 2003-04 to 2005-06 in terms of the content assessed. In 2003-04, 99% of Edward's items assessed content in just two Standards – Algebra and Number. Although data was assessed only with 5% of the items in 2005-06, the remaining items were distributed more evenly among the other four standards. In 2003-04, none of the items assessed geometry content. In 2005-06, 27.1% of the items assessed geometry content. In addition, the items assessed more content categories at the eighth grade in 2005-06. In 2003-04 Edward's items assessed only 5 of the 15 eighth-grade content categories; items in 2005-06 assessed 10 of the 15. While items were more evenly distributed among the five standards and the 15 content categories at the eighth grade, the same percentage of his items assessed eighth-grade content. In 2003-04, 16% of his items assessed content below the eighth grade and 31.5% assessed content above the eighth grade (47.5% total). In 2005-06, 35.6% of his

items were below eighth grade while 12.4% were above (48% total). While little more than half of the items assessed eighth-grade content each year, the remaining items changed from predominantly above eighth-grade items in 2003-04 to predominantly below eighth-grade items in 2005-06. This change is especially interesting since these items were given to advanced students taking a ninth-grade equivalent course.

Depth of Knowledge. Tables C10 and D10 report that over 90% of Edward's items were at level 1 during both the 2003-04 and 2005-06 school years. Edward used level 1 items 95.3% of the time during 2003-04. The remaining 4.7% of his items were level 2 items, with no level 3 items. Edward slightly increased the percentage of level 3 items to 0.6% by including one item in 2005-06. Other changes in terms of depth of knowledge were also small. Level 1 items accounted for 92.1% and level 2 accounted for 7.3% of his items in 2005-06. These large percentages of level 1 items were much higher than the 61.8% of level 1 items released from the OAT. It is worth noting that this large percentage of level 1 items was administered to an advanced group of eighth-grade students who were receiving ninth-grade credit.

Summary of Interview and Assessment Data. Edward was correct in predicting that his assessments changed from 2003-04 to 2005-06. In 2003-04, 99% of his items assessed either number or algebra. In 2005-06 his items were more evenly distributed among the standards. Much of this change was related to the change from an algebra course in 2003-04 to an integrated mathematics course in 2005-06. This change included increasing the percentage of items assessing geometry, from 0% in 2003-04 to 27.1% in 2005-06. Edward also recognized the value the textbook materials placed on algebra and geometry, and he supplemented items for the other three standards. Edward also assessed

more of the eighth-grade content categories increasing from just 5 of the 15 in 2003-04 to 10 of 15 in 2005-06. However, the percentage of items assessing content below the eighth grade increased from 16% in 2003-04 to 35.6% in 2005-06. This finding reflects the conflict that Edward discussed in teaching both eighth- and ninth-grade content. The analysis revealed that, despite teaching the equivalent of a ninth-grade course, over one-third of the items assessed content below the eighth grade in 2005-06. In terms of depth, over 90% of Edward's items were level 1 items in both 2003-04 and 2005-06. This finding is surprising considering that Edward submitted tests from his advanced eighth-grade class. Although Edward claimed to revise textbook items to include more writing and explanation, and despite that these items were for advanced students, the vast majority of his items were still at level 1.

Linda

Linda was in her 13th year of teaching during the 2005-06 school year. She held a grades 1-8 license to teach all subjects, and she submitted tests from her regular eighth-grade mathematics class. Linda used the CMP materials during both 2003-04 and 2005-06 school years. She also generally used the tests provided with these materials; however she made adjustments to some items as she deemed necessary. In deciding what mathematics content to test, she considered the eighth-grade indicators, the mathematics content covered by teachers in grades six and seven, and where her students needed to be as ninth graders. In terms of difficulty, she tried not to make items too difficult because she wanted students to have a chance to correctly answer each item. She stated that the primary purpose of tests was to allow students to show their work and reveal conceptual development. She indicated that her tests assessed content that students should know and

remember forever, and she expected students to do well on all of her tests. In terms of format, she used a CMP format that illustrated students' thinking. Linda said that ideally tests should serve as a re-teaching tool. She added, however, that in reality tests served as an end-of-learning tool to measure what students learned and how hard they had worked throughout the unit. While in some ways she acknowledged tests could be used as sorting tools, she wanted to use tests to inform her teaching. Other assessments that Linda used included class-work, observations, homework, and student interviews. She began student interviews in 2005-06 and struggled with how to transform information from the interviews into a grade. Linda's tests accounted for 25-33% of students' final grades. She attended OMAP training and OCTM sessions specific to the OAT between 2003-04 and 2005-06. She stated that these experiences helped her understand some of the details for the OAT, and they confirmed some things she already expected from the OAT. Linda also predicted that her assessments would be different in 2005-06 than in 2003-04. Because of how her district implemented CMP, she taught fewer seventh-grade modules in 2005-06 than in 2003-04 and was able to teach more eighth-grade modules.

Mathematics Content. Table C7 reports that Linda's 2003-04 assessments included 35.5% of the items in algebra, almost 31% in number, 22.2% in measurement, and almost 11% in data. None of her items assessed geometry content. Tables C4-C6 reveal that her most assessed content categories were *Use Algebraic Representations* with 32.7% and *Number Below Eighth Grade* with 20%. Therefore, one of every five items assessed number content below the eighth grade. Linda's items in 2003-04 assessed 9 of the 15 eighth-grade content categories.

Tables D4-D6 reveal that Linda's items assessed 12 of the 15 eighth-grade content categories in 2005-06. *Use Algebraic Representations* and *Number Below Eighth Grade* were her most assessed content categories with 29.9% and 10.3%, respectively. Table D8 reports that Linda assessed content considered below the eighth-grade level with 27.4% of her items in 2005-06. Table D7 reports that algebra was the most assessed standard with 32.5%, while data was the least assessed with 6.8%. Remaining items were evenly distributed among the other three standards, with 23.9% in number, 18.8% in geometry, and 17.9% in measurement.

Algebra was Linda's most assessed standard during both 2003-04 and 2005-06, with approximately one-third of the items. *Use Algebraic Representations* was also her most assessed categories both years. In 2003-04 none of Linda's items assessed geometry content. In 2005-06, 18.8% of her items assessed geometry content. Linda also increased from assessing 9 of the 15 eighth-grade content categories in 2003-04 to assessing 12 of the 15 in 2005-06. The three additional categories were all geometry categories. Linda also showed a small increase in the percentage of items that assessed content below the eighth grade, going from 22.7% in 2003-04 to 27.4% in 2005-06.

Depth of Knowledge. Table C10 reports that Linda used 87.3% of her items to assess level 1 knowledge in 2003-04, 11.3% to assess level 2 knowledge, and 0.9% to assess level 3 knowledge. Table D10 reports that the percentage of level 3 items did not change in 2005-06. However, Linda decreased the percentage of level 1 items to 79.5% and increased the percentage of level 2 items to 19.7%. These changes in levels 1 and 2 were significant compared to other teachers in the group. While Linda still used level 1 items almost 80% of the time in 2005-06, and this was still a larger percentage than the

Table 9

Linda's Assessment Data, 2003-04: Mathematics Content and Depth of Knowledge

N = 110

	Depth of Knowledge			
Standards	Level 1	Level 2	Level 3	Totals
Number	30.0%	0.9%	0.0%	30.9%
Measurement	20.0%	2.7%	0.0%	22.7%
Geometry	0.0%	0.0%	0.0%	0.0%
Algebra	31.8%	3.6%	0.0%	35.5%
Data	5.5%	4.5%	0.9%	10.9%
Totals	87.3%	11.8%	0.9%	100.0%

Table 10

Linda's Assessment Data, 2005-06: Mathematics Content and Depth of Knowledge

N = 117

	Depth of Knowledge			
Standards	Level 1	Level 2	Level 3	Totals
Number	21.4%	2.6%	0.0%	23.9%
Measurement	16.2%	1.7%	0.0%	17.9%
Geometry	11.1%	7.7%	0.0%	18.8%
Algebra	29.1%	3.4%	0.0%	32.5%
Data	1.7%	4.3%	0.9%	6.8%
Totals	79.5%	19.7%	0.9%	100.0%

released OAT items (61.8%), the cognitive demand of her items increased from 2003-04 to 2005-06.

Content-Depth Intersections. Linda's assessment items showed significant changes between 2003-04 and 2005-06. These changes can be analyzed further by considering the content and depth of her assessments simultaneously. Tables 9 and 10 illustrate Linda's assessment data in terms of both mathematics content and depth of knowledge. As stated earlier, Linda included no items assessing geometry in 2003-04, but had 18.8% of her items as geometry in 2005-06. Table 10 shows that over one-third of her geometry items in 2005-06, 7.7% of all of her items, were level 2 items. Also noted earlier, Linda's percentage of level 1 items decreased from 87.3% to 79.5%, and level 2 items increased from 11.8% to 19.7%. In addition to a significant proportion of level 2 items in the Geometry Standard, well over half of Linda's data items in 2005-06 were level 2. She was only one of two participants to have any standard assessed with a larger percentage of level 2 items than level 1 items.

Summary of Interview and Assessment Data. In the interview, Linda predicted that the items changed from 2003-04 to 2005-06, and the assessment data confirm this prediction. Linda changed from having no geometry items in 2003-04 to having 18.8% of the items assess geometry in 2005-06. This could be a result of teaching more of the eighth-grade modules. However, despite teaching fewer seventh-grade modules, the percentage of items that assessed content considered below the eighth-grade level actually increased from 22.7% in 2003-04 to 27.4% in 2005-06. Linda also described using test items such as those provided by CMP and items that provided insight into student thinking. This change seems to be related to the cognitive demand of Linda's

items. In 2003-04, 87.3% of her items were level 1, and this percentage decreased in 2005-06 to 79.5%. This percentage was the second lowest of level 1 items of the nine teachers in 2005-06.

Summary of 2003-04 Assessment Data: Mathematics Content

The summary columns in Tables C1-C6, as well as the bottom rows of tables C7-C9 summarize the aggregated teacher assessment data from 2003-04 in terms of mathematics content. They are reported this way to aid in comparing percentages to indicators and released OAT items, which are also reported within these tables..

The most assessed mathematics content category in 2003-04 across the nine teachers was *Use Algebraic Representations*. The average percentage of items in this category was 28.6%. This is consistent with the finding that *Use Algebraic Representations* also had the highest percentage of indicators at eighth grade with 19.6%. *Number Below Eighth Grade* was assessed with an average of 14.3% of teachers' items and was the only other category that averaged more than 10% across all teachers' items. This content obviously received much stronger attention from teachers than recommended by Ohio's ACS, as no content considered below the eighth grade appeared in the eighth-grade indicators. *Use Measurement Techniques and Tools* and *Statistical Methods* received much less attention from teachers' assessments than the emphasis found in the eighth-grade indicators. *Use Measurement Techniques and Tools* included 15.7% of the indicators; yet teachers averaged assessing this category with only 4.4% of the items. Likewise, *Statistical Methods* included 11.8% of the eighth-grade indicators, but teachers averaged assessing this category with only 1.6% of the items.

Teachers focused their assessment items most heavily in the Algebra Standard with an average of 47.4% of teachers' items assessing content in this standard. Number averaged nearly 25% of the assessment items, while measurement, geometry, and data each accounted for roughly 10% of the items. Although algebra had the largest percentage of eighth-grade indicators, teachers' assessments focused even more heavily on this standard. Only 26.3% of the eighth-grade indicators were algebra. Number also was emphasized slightly more by teachers (24.2%) compared to its percentage of eighth-grade indicators (15.7%), while measurement, geometry, and data each received less attention than what was given in the ACS.

Teachers' assessment items in 2003-04 illustrated more emphasis on content below the eighth grade. As stated earlier, *Number Below Eighth Grade* was the second most common content that teachers assessed in 2003-04, with an average of 14.3% of their items. Table C8 reveals that an average of 28.5% across all nine teachers' assessment items assessed content considered below the eighth grade. Only Frank's assessments revealed a small percentage of items below grade level with 3%. In addition, teachers in this sample averaged assessing 9.1 of the 15 eighth-grade content categories, and one of those categories, *Spatial Relationships*, was assessed by none of the teachers in this group during the 2003-04 school year.

Summary of 2005-06 Assessment Data: Mathematics Content

The summary columns in Tables D1-D6, as well as the bottom rows of tables D7-D9 summarize the aggregated teacher assessment data from 2005-06 in terms of mathematics content. They are reported this way to aid in comparing percentages to indicators and released OAT items, which are also reported within these tables.

The most assessed category from the teachers' assessments from 2005-06 was *Use Algebraic Representations*, with an average of 25.3% of the items across all teachers assessing content in this category. This was also the category with the largest percentages of standards and released items from the OAT. Teachers used a slightly larger percentage of assessment items in this category than both the percentage of eighth-grade indicators in this category (19.6%) and percentage of released items from the OAT (21.1%). *Number Below Eighth Grade* was the second most assessed category with an average of 14.7% of teachers' items. Neither the indicators nor OAT released items assessed content in this category. *Use Measurement Techniques and Tools* and *Statistical Methods* had smaller percentages of items than comparable percentages of the indicators at eighth grade. *Use Measurement Techniques and Tools* included 15.7% of the indicators, while teachers averaged assessing this content with only 3.4% of their assessment items. This content category was also assessed by 18.4% of the OAT released items. *Statistical Methods* included 11.8% of the eighth-grade indicators, while teachers averaged assessing this content with only 2.4% of their assessment items. This content category was also assessed by 6.6% of the OAT released items.

The Algebra Standard was assessed most often in 2005-06, as it was assessed with an average of 42.9% of teachers' assessment items. This standard had the highest percentage of eighth-grade indicators (31.4%) and released OAT items (26.3%); however, teachers placed more emphasis on this standard than both indicators and released OAT items. The Measurement and Data Standards had smaller percentages of items from teachers than percentages of indicators and released OAT items. Measurement accounted for 19.6% of the indicators at eighth-grade and was assessed

with 19.7% of the OAT released items. However, teachers only assessed this standard with an average of 9.4% of the items. Likewise, data accounted for 21.6% of the indicators at eighth grade and was assessed with 21.1% of the OAT released items. However, teachers assessed this standard with an average of 11.6% of the items. While algebra was a focus of the indicators at eighth grade and released OAT items, teachers placed much more attention on this Standard, with decreased attention to measurement and data.

Teachers' assessment items in 2005-06 focused heavily on content below eighth grade. As stated earlier, *Number Below Eighth Grade* was the second most assessed content category across the sample with an average of 14.7% of items. Table D8 illustrates that an average of almost one-third (32.5%) of teachers' items assessed content below eighth grade. Only Frank and Evelyn included less than 20% of items assessing content below the eighth grade. Also, teachers' assessments in 2005-06 averaged assessing 11 of the 15 eighth-grade content categories.

Summary of Assessment Data: Content Changes from 2003-04 to 2005-06

When considering the average percentage of items across the group of nine teachers in each of the content categories, no major changes in the content that teachers' items assessed between 2003-04 and 2005-06 existed. The most assessed categories during 2003-04 were *Use Algebraic Representations* with an average of 28.6% and *Number Below Eighth Grade* with an average of 14.3%. These two categories were also the most assessed in 2005-06 with 25.3% and 14.7%, respectively. Therefore an average of 40% of teachers assessment items focused on just two content categories. While *Use Algebraic Representations* was also a focus of the indicators at eighth grade and released

items from the OAT, teachers gave this category more attention than either the ACS or the OAT. *Number Below Eighth Grade* was heavily assessed by the teachers during both years; however, no eighth-grade indicators addressed content below the eighth grade and no released OAT items assessed number content below the eighth grade.

Teachers' assessments were similarly distributed among the five standards in both 2003-04 and 2005-06. Algebra was still the most assessed standard in 2005-06 with an average of 42.9%. This percentage decreased from 47.4% in 2003-04. However, these percentages are significantly higher than the percentages of indicators and OAT released items that fall under the Algebra Standard at the eighth grade. During both years teachers also included a higher percentage of number items compared to the percentages of indicators and OAT released items. They also included a smaller percentage of measurement and data items compared to the percentages of indicators and OAT released items. Geometry had the largest increase in percentage of items, making up just 8.8% of the items in 2003-04 and 13.4% in 2005-06. According to tables C7 and D7, this increase was due largely to Edward and Linda. Neither teacher included any items assessing geometry in 2003-04. In 2005-06 Edward included 27.1% and Linda included 18.8% assessing geometry content.

An examination of the content categories within these standards revealed that teachers' items assessed more content at the eighth-grade level in 2005-06. In 2003-04, teachers averaged assessing 9.1 of the 15 eighth-grade content categories. In 2005-06, they averaged assessing 11 of the 15 eighth-grade content categories. From 2003-04 to 2005-06, teachers in this sample averaged assessing nearly two more content categories found at the eighth grade.

This finding seems to contradict another finding from these data. Tables C8 and D8 reveal that the percentage of items assessing content considered below the eighth grade increased from 2003-04 to 2005-06. In 2003-04, teachers averaged 28.5% assessing content below the eighth grade. In 2005-06, that percentage increased to 32.5%. Frank, Helen, and Edward showed the greatest increase in the percentage of items that assessed content below grade level. Edward's results were especially interesting since he submitted tests from an advanced class that was equivalent to a ninth-grade course. Despite the presence of the ACS since 2002 and released OAT items since January of 2005, teachers in this group increased the percentage of items that assessed content considered below eighth grade. Although the items in 2005-06 were more evenly distributed among the 15 eighth-grade content categories, teachers had a lower percentage of items assessing the recommended eighth-grade content.

As noted earlier, Edward and Linda made significant changes in the percentage of geometry items from 2003-04 to 2005-06. However, a large percentage of these items assessed content below eighth-grade level. Table D5 reveals that 14.1% of Edward's items in 2005-06 assessed geometry content considered below grade level. That is, over half of the 27.1% of his items that fell under geometry in Table D7. Similarly, 7.7% of Linda's items assessed geometry content below grade level, which is over one-third of the 18.8% of her items that fell under geometry. Despite having items more evenly distributed over the five standards, a large percentage of items assessed content considered below the eighth grade.

Summary of Assessment Data: Items with Multiple Content Categories

A few items were coded into multiple content categories. Many items were divided into discrete parts. As such, these items were coded as separate items. For example, items that had explicitly listed parts (a) – (d) were coded as four separate items. Tables 11 and 12 list the number of items that were coded under multiple content-categories in 2003-04 and 2005-06, respectively. The tables also indicate the specific content categories to which each item was coded, as well as the depth level of each of the items.

The purpose of examining these items was to determine if teachers assessed the indicators in ways consistent with the vision for school mathematics articulated in the NCTM (2000) and ODE (2002) Standards documents. The authors of both documents

Table 11

2003-04 Assessment Data: Multiple Content Items

Teachers	No. of items	Content Categories	Depth Level
Sam	1	Use Patterns, Relations, and Functions Use Algebraic Representations	2
Wanda	2	Measurement Units Use Patterns, Relations, and Functions	1
	1	Data Below Eighth Grade Data Above Eighth Grade	3
Evelyn	1	Use Patterns, Relations, and Functions, Use Measurement Techniques and Tools	1

Table 12

2005-06 Assessment Data: Multiple Content Items

Teachers	No. of items	Content Categories	Depth Level
Wanda	1	Measurement Units Use Algebraic Representations	2
Evelyn	3	Use Algebraic Representations Use Measurement Techniques and Tools	2
Henry	1	Number Below Eighth Grade Measurement Below Eighth Grade	1
Edward	5	Use Patterns, Relations, and Functions Geometry Below Eighth Grade	1
OAT	1	Use Algebraic Representations Use Measurement Techniques and Tools	1
	1	Use Algebraic Representations Use Patterns, Relations, and Functions Analyze Change	2
	1	Measurement Units Use Algebraic Representations	1
	1	Data Collection Use Algebraic Representations	3
	1	Characteristics and Properties Spatial Relationships	2

recommend that the Standards not be taught in isolation of one another, but as integrated topics. Tables 11 and 12 show that some of these items were coded under two content categories. However, this finding reveals the overlap in the grade level indicators more than teachers' ability to integrate multiple standards into test items. Four of the items, three from teacher assessments and one from the released OAT items, assessed content under the categories of *Use Algebraic Representations* and *Use Measurement Techniques and Tools*. Each of these items used formulas to convert measurements (e.g. Celsius to Fahrenheit) and was considered both a measurement item and an algebra item because of how the indicators were written, not because of teachers' ability to frame items that span multiple content categories.

Of the nearly 6000 items that were coded, only 19 were coded as assessing multiple content categories. Tables 11 and 12 also revealed that 5 of the 76 released OAT items were coded as assessing multiple content categories, including one item that assessed three different content categories. This OAT percentage is significantly higher than that of classroom teachers' items.

Summary of Assessment Data: Depth of Knowledge

Tables C10 and D10 show the percentages of each teachers' items that were categorized as levels 1, 2, and 3 for both 2003-04 and 2005-06. The last two rows report the mean percentage of items at each depth level for the group of teachers and the percentage of OAT released items at each depth level.

Like the OAT items, teachers' assessments during the 2003-04 and 2005-06 school years included large percentages of level 1 items. However, teachers' test items had much higher percentages of level 1 items than the released OAT items (61.8%). The

teachers in the sample averaged 86.6% of their items at level 1 in 2003-04, and 86.4% in 2005-06. The mean percentage of teachers' level 2 and level 3 items also changed very little from 2003-04 and 2005-06. The percentage of level 2 items increased from 11.3% in 2003-04 to 12.7% in 2005-06, and the percentage of level 3 items dropped from 2.1% to 1%. While the mean percentage of level 3 items decreased, more teachers used at least one level 3 item on their tests in 2005-06. In 2003-04, five teachers used no level 3 items, while only two teachers used no level 3 items in 2005-06. Other than Frank, teachers showed very little change in level 3 items because so few of the items were used.

Only one classroom teacher in the group used a smaller percentage of level 1 items than the percentage of released OAT items. Approximately 58% of Sam's items during the 2005-06 school year were level 1 items, compared to 61.8% of the released OAT items. Six of the nine teachers decreased the percentage of level 1 items that they used from 2003-04 to 2005-06. However, three of these teachers, Evelyn, Nancy, and Edward, still used over 90% level 1 items in 2005-06. Frank, Helen, and Henry each increased the percentage of level 1 items from 2003-04 to 2005-06. This increase for Helen and Henry caused them to use over 95% of their 2005-06 test items at level 1.

Evelyn and Edward submitted items from their advanced eighth-grade classes, which were ninth-grade equivalent mathematics courses. Both Evelyn and Edward decreased the percentage of level 1 items from 2003-04 to 2005-06. Almost 98% of Evelyn's items were level 1 in 2003-04, and this percentage decreased to approximately 93% in 2005-06. Approximately 95% of Edward's items were level 1 in 2003-04, and this percentage decreased to 92.1% in 2005-06. Despite teaching advanced classes,

Evelyn and Edward continued assessing students at very low depths of knowledge, as over 90% of their items continued to be level 1.

The assessments of Sam and Linda revealed a decrease in the number of level 1 items over the two-year period. They had the lowest percentage of level 1 items in the group of teachers in 2005-06. They also had the highest percentage of level 2 items across the nine teachers in 2005-06. Interestingly, they were the only teachers who used NSF-funded curricula. During the interviews, they reported that the majority of their assessment items were taken from these materials. In fact, seven of the teachers reported using many test items provided in their curriculum materials. Frank and Edward were the only teachers that did not. Frank's tests were developed by his curriculum leader; Edward's textbook materials did not come with tests, so he created his own. Sam and Linda, however, both using CMP materials, had the lowest percentage of level 1 items and highest percentage of level 2 items in 2005-06.

Summary of Assessment Data: Content and Depth Intersections

This section examines the intersections of mathematics content and depth of knowledge simultaneously for the aggregated test data from the nine teachers. Table 13 illustrates that teachers emphasized assessing algebra content at depth of knowledge level 1 during 2003-04. Teachers averaged assessing algebra with level 1 items with 41.4% of their items in 2003-04. A second emphasis was number items at level 1 with an average of 22.6% of teachers' items. These two cells accounted for 64% of teachers' assessment items in 2003-04. Earlier sections support this finding that level 1 items in algebra and number were emphasized. Teachers' most assessed standards in 2003-04 were algebra (47.4%) and number (24.2%), with an average of 86.6% of their items at level 1.

Table 13

Aggregated Assessment Data, 2003-04: Mathematics Content and Depth of Knowledge

	Depth of Knowledge			
Standards	Level 1	Level 2	Level 3	Totals
Number	22.6%	1.4%	0.1%	24.2%
Measurement	9.3%	1.3%	0.2%	10.8%
Geometry	7.2%	1.7%	0.0%	8.8%
Algebra	41.4%	4.9%	1.1%	47.4%
Data	6.2%	2.0%	0.7%	8.9%
Totals	86.6%	11.3%	2.1%	100.0%

Table 14

Aggregated Assessment Data, 2005-06: Mathematics Content and Depth of Knowledge

	Depth of Knowledge			
Teachers	Level 1	Level 2	Level 3	Totals
Number	22.1%	1.7%	0.0%	23.7%
Measurement	7.9%	1.5%	0.1%	9.4%
Geometry	11.4%	2.0%	0.0%	13.4%
Algebra	38.0%	4.4%	0.5%	42.9%
Data	7.3%	3.2%	0.4%	11.0%
Totals	86.4%	12.7%	1.0%	100.0%

Similar results were found with 2005-06 assessment data from teachers. Table 14 reveals that algebra and number, both at level 1, were the most emphasized cells. Algebra level 1 accounted for an average of 38% of the teachers' items in 2005-06, while number level 1 accounted for an average of 22.1%. While this finding represents a decrease from 2003-04, an average of 60.1% of teachers items still were in these two cells.

As mentioned earlier, the average percentage of teachers' items assessing the Geometry Standard increased from 8.8% in 2003-04 to 13.4% in 2005-06. Considering the intersection of geometry and depth, most of the increase occurred in level 1 items. Level 1 items in geometry increased from 7% in 2003-04 to 11.4% in 2005-06, while level 2 items only increased from 1.7% to 2.0%. In addition, teachers used no level 3 items to assess geometry in either year.

Tables 13 and 14 also reveal that data items also reflected a small increase from 8.9% in 2003-04 to 11% in 2005-06. This change was not large, but the depth level of data items changed more dramatically. In 2003-04, the ratio of level 2 items to level 1 items was 1:3, with 2% data level 2 items compared to 6.2% data level 1 items. In 2005-06 the ratio of level 2 to level 1 increased closer to 1:2, with 3.2% level 2 and 7.3% level 1. In examining classroom assessment data from the group of teachers, this standard was the only one in either year that had this high of a proportion of its items above level 1. All other standards had a much lower ratio of level 2 to level 1 items. The ratio of data items in teachers' assessment items is consistent with data items from the OAT. Released OAT items from data had a level 2 to level 1 ratio of 10.5% to 9.2%. Data was the only standard where more level 2 items were used than level 1 items. Both the OAT

items and teachers items had a higher percentage of level 2 items in data than any other standard.

Interview Data

Each teacher was interviewed for approximately 20-30 minutes. These interviews were conducted to reveal each teacher's purpose of testing and the process that teachers used in developing their classroom tests. This section presents a summary of the interview data, along with a summary of assessment item data, with regard to common themes.

Textbook Materials

Seven of the nine participants reported that they depended on their curriculum materials as the basis for their classroom tests. Frank and Edward were the only two who did not rely on their curriculum materials for test items. Frank's tests were developed by the curriculum leader in his district, while Edward created his own tests because of the limited resources provided with his curriculum materials. The other seven participants used tests provided by their textbook publishers and reported that these materials played a role in deciding how they developed their eighth-grade tests. The nine teachers averaged 28.5% of items assessing content below grade level in 2003-04, and 32.5% in 2005-06. Frank and Edward averaged 9.5% in 2003-04 and 26.4% in 2003-04. The seven participants depending on their textbooks averaged 33.9% in 2003-04 and 34.2% in 2005-06. The data reveal that the teachers relying on their textbook had higher percentages of items assessing content below the eighth-grade than those not relying on their textbooks.

In terms of depth of knowledge, the nine teachers averaged 86.6% of their items as level 1 in 2003-04, and 86.4% in 2005-06. Frank and Edward had similar percentages

to the seven relying on their curriculum materials. They averaged 79.9% level 1 in 2003-04 and 87.5% in 2005-06. The seven teachers averaged approximately the same percentage of level 1 items in both years, 88.5% in 2003-04 and 86.0% in 2005-06. Depending on the textbook for test items resulted in over one-third of the items assessing content below grade level and over 85% of the items assessing content at depth of knowledge level 1.

Of the seven that depended on their textbooks, two teachers used NSF-funded curricula. Sam and Linda used CMP materials with their eighth-grade students, and the tests from these materials served as a basis for their items. Teachers using these items assessed more content at the eighth grade than teachers using other textbook items. The average percentage of items assessing content below the eighth grade for the seven participants using their textbooks was approximately 34% for both 2003-04 and 2005-06. Sam and Linda's percentages were both below these averages, with Sam at 28.9% in 2003-04 and 20.9% in 2005-06 and Linda at 22.7% in 2003-04 and 27.4% in 2005-06. In addition, Sam and Linda had smaller percentages of level 1 items than the other teachers using their textbook materials as a basis for their test items. These seven averaged 88.5% at level 1 in 2003-04, while Sam's percentage was 61.9% and Linda's percentage was 87.3%. The seven averaged 86% at level 1 in 2005-06, while Sam and Linda had the lowest percentage of level 1 items in 2005-06 at 58.2% and 79.5%, respectively.

Eighth-Grade Indicators

Six of the nine teachers mentioned either the Ohio ACS or specifically the eighth-grade indicators as having an impact on the mathematics content that they assessed. Despite reporting that they were attentive to the Ohio ACS, these participants averaged

similar percentages of items that assessed content below the eighth grade in both 2003-04 and 2005-06 compared to those who were not as attentive to the ACS. In 2003-04, the six who were attentive to the ACS averaged 27.6% of items assessing content considered below the eighth-grade level, while the remaining three averaged 30.3%. In 2005-06 the six averaged 32.5% while the remaining three averaged 32.4%. Despite claiming to focus on Ohio's Standards to determine the mathematics content of their tests, these six teachers averaged nearly one-third of their test items assessing content below the eighth grade.

Advanced Classes

Two of the teachers, Evelyn and Edward, submitted test items from their advanced classes. Evelyn had lower percentages of items (13.1% in 2003-04 and 13.2% in 2005-06) that assessed content below the eighth grade. With a different pattern, Edward used only 16% items that assessed content below the eighth grade in 2003-04 and 35.6% in 2005-06. Other factors besides the OAT could have impacted this change. Edward's advanced course changed from an algebra course in 2003-04 to an integrated mathematics course in 2005-06. He also used new curriculum materials as part of this new course. This increase in the percentage of items assessing content below the eighth grade is still a significant increase, especially considering that this class was the equivalent to a ninth-grade mathematics course.

Both Evelyn and Edward used over 95% of the items at level 1 in 2003-04. While both decreased the percentage of level 1 items in 2005-06, they still used over 90% of the items as level 1. In the interview Evelyn discussed wanting to challenge all her students and stated she had incorporated varying degrees of difficulty in her test items. She

disliked using multiple-choice items because they could not be "in-depth" enough for her. Test grades accounted for approximately 50% of students' grades in Evelyn's class. Edward stated that his tests were easier than his homework items, where students were challenged the most. He avoided using multiple-choice items as well in order to make his test items more challenging. Between 40-50% of a student's final grade was determined by test grades in Edward's class. Given that Evelyn and Edward's test items were designed for advanced students and that tests accounted for a larger percentage of students' grades, their items included a surprisingly high percentage of level 1 items.

Recognizing Change

Five of the nine teachers (Henry, Sam, Linda, Edward, and Frank) predicted that their assessments had changed between 2003-04 and 2005-06. For Henry, Sam, Linda, and Edward, the assessment data largely confirmed their predictions. Henry stated that only the number of tests he gave changed, and the data confirmed that he gave 10% fewer items in 2005-06 compared to 2003-04. In addition, his items changed very little with respect to content and depth, which was consistent with his prediction. Sam and Linda both stated that their items would change in terms of the mathematics content. The item analysis data confirmed this, as Sam significantly increased the percentage of items that assessed data, and Linda significantly increased the percentage of items that assessed geometry. Edward discussed changing his curriculum materials in 2005-06 to assess more of the standards. The item analysis data confirm this finding, as in 2005-06 items were more evenly distributed among the five standards, and in 2003-04 99% of items assessed number and algebra.

Edward, along with Frank, predicted that the items addressed the indicators and OAT better in 2005-06. In this regard, the data revealed this prediction to be incorrect for both. The data for Edward and Frank revealed an increase in the percentage of items that assessed content below the eighth grade. The items on Edward's assessments changed from 16% in 2003-04 to 35.6% in 2005-06. The items on Frank's assessments changed from 3% in 2003-04 to 17.1% in 2005-06. In terms of content, both Edward and Frank used more items assessing content below the eighth grade in 2005-06. In addition, Frank increased the percentage of level 1 items from 64.4% in 2003-04 to 82.9% in 2005-06. Edward also had a high percentage of level 1 items, with 95.3% in 2003-04 and 92.1% in 2005-06. The released OAT items used a much smaller percentage of level 1 items (61.8%). For Frank and Edward, having items that better addressed the indicators and OAT resulted in using more items that assessed content below the eighth grade and a large percentage of items that were recall or procedural.

Professional Development

Four of the nine teachers reported receiving professional development specific to assessment between 2003-04 and 2005-06. Two teachers, Wanda and Linda, attended OMAP training. Wanda stated that the training was not helpful, although she decreased the percentages of level 1 items from 2003-04 to 2005-06. Linda reported that the OMAP training, as well as OCTM sessions on the OAT, confirmed what she already expected from the OAT. Linda also decreased the percentage of level 1 items in 2005-06, and her items were more evenly distributed among the five standards. Nancy received training on short-cycle assessments in her district. Between 2003-04 and 2005-06, she added three short-cycle assessments to her tests but did not change the other tests

that she administered. Because the other tests made up a larger percentage of her items, her test data did not change dramatically from 2003-04 to 2005-06. Henry received training specific to using data from the OAT. He and other teachers were given time and instruction on how to analyze the data and make decisions about what to cover better in the future. However, Henry's items changed very little from 2003-04. The only noticeable change was that his most assessed category in 2005-06 was algebra content above the eighth grade with over 22%.

Overall, this group of teachers did not receive significant training specific to assessment. Some did attend workshops and sessions specific to the OAT or indicators. However, these teachers stated that the sessions were not helpful or that the sessions merely confirmed what they already knew about the OAT or the indicators.

CHAPTER V

CONCLUSIONS

With the passage of the Elementary and Secondary Education Act (ESEA) in 2002, all states implemented accountability systems to document student achievement. States are required to develop academic content standards in the hopes that these rigorous standards will drive improved classroom curriculum and instruction and ultimately improve student performance. States are also required to give annual assessments to measure the degree to which students are meeting these standards (Lonergan, 2003; United States Department of Education (USDE), 2005). In March of 2006 in Ohio, achievement tests were given to all students in grades 3-8 in reading and mathematics.

Standardized testing is not new to education in the United States. In the past, achievement tests have been used to measure students' progress, sort students by ability, and track students into various career paths (Gallagher, 2003). Supporters of the recent standards movement, and ensuing testing movement, believe that standards and tests provide educators a common focus (Fremer & Wall, 2003; Popham, 1987; Popham, 2003). Some tests have been developed to encourage the type of curriculum and instruction that professional organizations such as the National Council of Teachers of Mathematics (NCTM) have promoted for years (Cohen & Ball, 1990; Firestone, Mayrowetz, & Fairman, 1998; Schorr, Firestone, & Monfils, 2003; Wilson, 2003).

Some believe that testing negatively affects students' learning (Bracey, 1987). Traditionally, tests tended to assess only things that can easily be measured, rather than students' creativity, critical thinking, and persistence (Bracey, 1987). Previous studies have shown that tests encourage teachers to (1) narrowly focus only on the content tested, (2) increase the amount of time teaching students test-taking skills, and (3) spend less time on good instruction of the content (Adams, Pedulla, & Madaus, 2003; Taylor, Shepard, Kinner, & Rosenthal, 2002). This study investigated the effects that recent mathematics tests in Ohio have had on teachers' classroom mathematics assessments.

Influence on Teacher Practices

Recent accountability systems and state tests were initiated to improve the education provided to students (Lonergan, 2005; USDE, 2002). At the center of this education are teachers and their classroom practices. While tests may be implemented by state agencies with the intention to reform and change teachers' classroom instructional and assessment practices, other factors influence these practices. This section highlights some of these factors and reports what previous research has revealed regarding the effects of state tests on classroom assessments.

Teachers' beliefs about what it means to know and do mathematics impact how they carry out the process of teaching and assessing mathematics in their classroom (Mewborn, 2002; Rousseau, 2004). Teachers with "traditional" beliefs about the nature of mathematics tend to focus on demonstrating commonly used procedures for students to mimic, followed by students practicing these procedures (Brown, Cooney, & Jones, 1990; Rousseau, 2004; Thompson, 1992). Teachers with "reform" beliefs about the nature of mathematics tend to engage students in mathematics through problem solving

and reasoning, as opposed to having students listen to and watch the teacher solve problems, reason, and communicate about mathematics (Thompson, 1992).

Teachers' beliefs about learning also influence their instructional practices (Rousseau, 2004). Teachers who believe that learning involves making content interesting and motivating use different instructional practices than teachers who believe that learning is achieved through hard work, discipline, and diligence (Joram & Gabriele's, 1998; Rousseau, 2004).

Research reveals that teachers' beliefs about teaching also impact their instructional practices (Mewborn, 2002; Pajares, 1992; Thompson, 1992). Pre-service teachers' beliefs about teaching are well established prior to studying to be a teacher (Ball, 1998; Pajares, 1992). Teaching is unlike other professions. Individuals studying to be doctors and lawyers do not have as much first-hand experience in these fields as those learning to become teachers. Pre-service teachers have spent much of their young life watching teachers do their work. Their beliefs about what it means to teach have evolved over long periods of time; therefore, the beliefs can be difficult to change (Lerman, 1997; Mewborn, 2002).

Teachers' knowledge of mathematics also influences their instructional practices (Ball 1988; Cooney, Badger, and Wilson, 1993; Ma, 1999, Shulman, 1986). How this knowledge is organized also determines how teachers structure their classroom instruction (Ma, 1999). Teachers with deep conceptual understandings of mathematics teach mathematics very differently than those with superficial understandings of mathematics (Ma, 1999).

Individual school settings and community cultures also influence instructional practices. Building principals and collegial relationships among teachers can also influence instructional practices (Hoy & Hoy, 2003; Louis, Kruse, & Marks, 1996; Rousseau, 2004; Taylor, 2004). Collaborations among teachers can also change what mathematics content is taught, as well as the manner in which it is taught (Rousseau, 2004; Taylor, 2004). Finally, parents can influence instructional practices. Wilson (2003) cited examples whereby parents and community groups challenged the way mathematics was taught in classrooms by pressuring teachers to make changes in their instructional practices. While these studies found that particular school settings influenced instructional practices, Stigler and Hiebert (1999) noted that teaching is largely a cultural activity. Their analysis of classrooms across different countries revealed that United States teachers generally follow a very similar script, regardless of school settings.

Local, state, and federal policies also influence instructional practices (Cohen & Ball, 1990; ODE, 2002; Wilson, 2003). In the last fifty years, policies regarding education were enacted in response to (1) the launching of Sputnik, (2) the report *A Nation at Risk*, (3) poor NAEP results, and (4) international studies such as The Third International Mathematics and Science Study (TIMSS) and The Program for International Student Assessment (PISA). The recent re-authorization of ESEA and the ensuing achievement tests, including the Ohio Achievement Test (OAT) in eighth-grade mathematics, provide more recent examples of policies intended to influence teachers and their instructional practices.

Testing Influence on Curriculum, Instruction, and Assessment

This section highlights previous research on the influence state tests have had on curriculum, instruction, and assessment.

Curriculum

Recent studies have shown that the presence of state tests resulted in teachers narrowing the content that they teach (Abrams, Pedulla, & Madaus, 2003; Firestone, Mayrowetz, & Fairman, 1998; Taylor, Shepard, Kinner, & Rosenthal, 2002). Teachers who taught multiple content areas directed more attention to those areas tested and less attention to areas not tested (Taylor, Shepard, Kinner, & Rosenthal, 2002). Teachers who worked in only one content area also narrowed their focus by allocating more time on content they knew would be tested and less time on content that would not be tested (Abrams, Pedulla, & Madaus, 2003; Firestone, Mayrowetz, & Fairman, 1998).

Instruction

Researchers (Abrams, Pedulla, & Madaus, 2003; Corbett & Wilson, 1991; Glassnapp, Poggio, & Miller, 1991; Taylor, Shepard, Kinner, & Rosenthal, 2002) also found that teachers react to tests in ways that contradict appropriate educational practices. In these studies, teachers tended to spend more time (a) teaching to the test in a game-like manner, (b) focusing on test-taking skills such as drills, (c) coaching for the test, and (d) practicing sample test items. Kupperminz, Shepard, and Linn (2001) reported that, while these types of activities improved test scores, scores did not reflect improved learning outside of the narrow focus of the test. Others have expressed concern that teachers have become deskilled as a result of the increased emphasis on standardized testing (Shepard, 2000; Smith, 1991). Teachers tend to gravitate to one "correct" way to teach, that is

"reduce a task to simpler components and drill it repeatedly until pupils have mastered it" (Smith, 1991, p. 11). Furthermore, multiple choice testing tends to lead to "multiple choice" teaching, which lessens the likelihood that teachers will assist students in developing conceptual understanding of the content (Smith, 1991).

Assessment

Abrams, Pedulla, and Madaus (2003) and McMillan, Myran, and Workman (1999) found that state tests influenced teachers' assessment practices. Teachers across the United States developed classroom assessments that mirrored the format of the state tests (Abrams, Pedulla, & Madaus, 2003). Black & Wiliam (1998) commented that standardized test items do not provide teachers with the insight into student learning and are poor examples for classroom assessment items. Cooney, Badger, and Wilson (1993) found that few teachers could produce assessment items that provided insight into student understanding. Not only did teachers lack the necessary skills for writing good assessment items, but state tests provided teachers poor examples of classroom assessment items.

Shepard (2000) argued that assessment practices are difficult to change. Perhaps change is difficult because assessments reflect teachers' beliefs about what is important (Cooney, Badger, & Wilson, 1993; NCTM, 2000), and changing teachers' beliefs is difficult (Lerman, 1997; Mewborn, 2002). Very few studies have examined the effects that state tests have on teachers' classroom assessment practices. No studies that examined these effects in terms of the specific mathematics content and the depth of knowledge that teachers assess seem to exist. In addition, those studies that examined the effects of state tests relied heavily on teacher self-reporting data. These results are

suspect, however, because research has shown that teachers' perceptions of their instructional practices often do not align with observational data (Cohen & Ball, 1990; Firestone, Mayrowetz, & Fariman, 1998; Schorr, Firestone, & Monfils, 2003).

Assessing Mathematics Content and Depth of Knowledge

Through its standards documents, NCTM set ambitious goals for the mathematics content that K-12 students should learn (NCTM, 1989; NCTM, 2000). ODE's standards also call for a focus on mathematics content from all five NCTM content standards: Number, Measurement, Geometry, Algebra, and Data. For example, the Ohio Academic Content Standards (ACS) at eighth grade list approximately the same percentage of indicators in each of these five standards. In addition to the mathematics content that students are expected to learn, NCTM advocates assessing students' full "mathematical power," not just their ability to perform routine procedures and isolated skills (NCTM, 1995). Through their work with the QUASAR Project, Smith and Stein (1998) emphasized the importance of exposing students to high demand assessment items in the classroom. Smith and Stein (1998) and Webb (1999) have developed frameworks for examining and assessment tasks at greater depth of knowledge levels, such as conceptual understanding and strategic thinking. Cooney, Badger, and Wilson (1993) found, however, that teachers were often unable to write items that assess students beyond memorization or performance of routine procedures.

Research Questions

With the new Ohio state tests in mathematics, administrators and others need to know whether teachers assess the content recommended by the Ohio Department of Education (ODE) in its Academic Content Standards (ACS). They also need to know

whether teachers assess students' full mathematical power through a variety of knowledge levels recommended by the ODE. Addressing this need, this study sought to answer the following questions:

4. How have eighth-grade teachers' assessments changed over the past two years relative to the mathematics content that they assess?
5. How have eighth-grade teachers' assessments changed over the past two years relative to the depth of knowledge that they assess?
6. What factors do teachers attribute to changes that may exist in their classroom assessment practices over the last two years?

Methodology

This study investigated the effects that the new eighth-grade Ohio Achievement Tests (OAT) had on eighth-grade mathematics teachers' classroom assessments over the last two years. Specifically, teachers' assessments used during the 2003-04 and 2005-06 school years were analyzed with respect to mathematics content and depth of knowledge. The eighth-grade mathematics OAT was first administered in March 2005. Therefore, during the 2003-04 school year, teachers had little knowledge of the test. Classroom tests were collected from eighth-grade teachers in both the 2003-04 and 2005-06 school years. Collecting teachers' classroom tests eliminated some of the issues associated with self-reporting that other studies have encountered.

Participants

Nine middle school mathematics teachers participated in the study. Sam was in his fifteenth year of teaching during 2005-06, and he held a grades 7-12 license to teach mathematics. Wanda was in her sixth year of teaching and held a grades 1-8 license.

Frank was in his third year of teaching with a grades 4-9 license to teach mathematics. Evelyn was in her 30th year of teaching and held both a grades 7-12 license and grades 1-8 license. Nancy was in her 35th year of teaching in 2005-06 and held a grades 1-8 license. Helen was in her 8th year of teaching and held a grades 1-8 license. Henry was in his 17th year of teaching and held a grades 7-12 license to teach mathematics. Edward was in his 9th year of teaching and held a grades 4-9 license to teach mathematics. Linda was in her 13th year of teaching and held a grades 1-8 license. Evelyn and Edward both submitted tests from their advanced classes, which were equivalent to ninth-grade courses. The other seven teachers submitted tests from their regular eighth-grade mathematics classes.

Test Analysis

All test items from each teacher were analyzed by the researcher, and each test item was coded in two ways. First, items were coded with regard to the mathematics content assessed. The 15 sub-standards for the eighth-grade indicators in the Ohio's ACS were used as mathematics content categories. In addition, items that did not assess eighth-grade content were coded as either above or below eighth grade, depending on whether the content assessed was found in grade-level indicators above or below the eighth grade. Second, each item was coded in terms of the depth of knowledge that it assessed. Webb's (1999) framework for analyzing mathematics items was used in this analysis. Webb identified four categories for classifying depth of knowledge: recall, skills/concepts, strategic thinking, and extended thinking.

Level 1 of Webb's (1999) framework is labeled *recall*. It includes items that elicit a rote response from students, such as recalling a basic fact or performing a simple

algorithm. Level 2 is labeled *skill/concept*. Level 2 items require mental processing beyond a habitual response and focus on student understanding of mathematical concepts and procedures. While level 1 has students follow a set procedure, level 2 items require students to make some decisions about how to approach the item. Level 3 is labeled *strategic thinking*. Level 3 items require planning and reasoning and a higher level of thinking than levels 1 and 2. Level 3 items have students draw conclusions, cite evidence, or develop a logical argument for concepts, and they often ask students to explain their reasoning. Level 4 is labeled *extended thinking* and items require complex reasoning and planning most likely over an extended period of time. They require applying significant conceptual understanding and higher-order thinking, and include tasks such as designing and conducting experiments and synthesizing ideas into new concepts. While Webb identifies four levels, only three of them were needed for this analysis. The fourth level did not apply to any of the classroom test or released OAT items. Released test items from the eighth-grade Ohio Achievement Test (OAT) also were coded for mathematics content and depth of knowledge.

Because comparisons were made between 2003-04 and 2005-06, teachers in the study must have taught the same level of mathematics in those two years. In addition, teachers needed to have access to the specific tests they administered in 2003-04. To gain insight into teachers' assessment practices and the results of the testing data, teachers participated in a 20-30 minute interviews with regard to their testing practices.

Findings

The analysis of items and interviews revealed interesting findings about the mathematics content and depth of knowledge of the teachers' classroom assessments. This section reports the findings of these analyses.

Mathematics Content

The item analysis revealed that a significant percentage of teachers' items assessed content considered below the eighth grade. In 2003-04, prior to the state test in eighth grade, teachers averaged 28.5% of their items below eighth grade. In 2005-06, since the presence of a state test, this average rose to 32.5%. While state tests intend to focus teachers' attention on the standards (Fremer & Wall, 2003; Popham, 1987; Popham, 2003), teachers in this study assessed more mathematics content considered below the eighth grade in the second year. While all five of the standards included items that assessed content below eighth grade, items focusing on number content were most common below the eighth grade. In fact, *Number Below Eighth Grade* was the second most assessed content category during both 2003-04 and 2005-06. Teachers assessed this category with an average over 14% of their items both years. In addition, the percentage of *Geometry Below Eighth Grade* items doubled from 2003-04 to 2005-06, increasing from 3.9% to 7.8%. All geometry items were included more often in 2005-06, but most of the increase resulted from items assessing geometry content below the eighth grade.

Teachers who were aware of the recommended eighth-grade content found in Ohio's ACS included significant percentages of items that assessed content below eighth grade. Six of the nine teachers reported that the Ohio ACS influenced the content that they assessed. These six teachers assessed content below the eighth grade with an

average 27.6% and 32.5% of their items in 2003-04 and 2005-06, respectively. The other three teachers, who did not report attending to the standards, averaged 30.3% and 32.4% in 2003-04 and 2005-06, respectively. Teacher's attention to and interpretation of the standards did not translate accurately into assessment items for those standards. Despite knowing the eighth-grade indicators, nearly one-third of teachers' items assessed content below eighth grade.

The analysis also revealed that the Algebra and Number Standards were over-emphasized by teachers in their assessment items compared to the emphasis found in the indicators and released OAT items. Teachers averaged assessing these two standards with approximately two-thirds of their items, while approximately 45% the indicators and released OAT items emphasized these standards. The released OAT items assessed algebra most often (26.3%), and algebra included the highest percentage of eighth-grade indicators (31.4%). Teachers assessed algebra content with an average of 47.4% in 2003-04 and 42.9% in 2005-06. Number was assessed with 18.4% of the released OAT items, and it included 15.7% of the eighth-grade indicators. Teachers' items assessed the Number Standard with almost 25% of their items in both 2003-04 and 2005-06.

The analysis also revealed that teachers' items at eighth grade were more evenly distributed among the eighth-grade content categories in 2005-06 than in 2003-04. From 2003-04 to 2005-06, teachers in this sample averaged assessing nearly two more content categories found at eighth grade. In 2003-04, teachers averaged assessing 9.1 of the 15 categories, and in 2005-06, they averaged assessing 11 of the 15. While a larger percentage of teachers' items assessed content below eighth-grade in 2005-06, the

remaining items were more evenly distributed among the 15 eighth-grade content categories.

Depth of Knowledge

The vast majority of teachers' assessment items focused on the lowest depth of knowledge level. Furthermore, from 2003-04 to 2005-06, the overall depth of knowledge that teachers assessed changed very little. On average, the nine teachers in the sample used level 1 items with over 86% of their items in both 2003-04 and 2005-06. This represents a large percentage of items that asked students to recall a basic fact or perform a routine procedure. Level 2 items accounted for an average of 11.3% in 2003-04 and 12.7% in 2005-06, while level 3 accounted for an average of 2.1% in 2003-04 and 1% in 2005-06. Although released items from the OAT also focused mostly on level 1 (61.8%), teachers assessed with a much larger percentage of level 1 items. The OAT released items also included 35.5% level 2 items and 2.6% level 3 items.

Six of the nine teachers decreased the percentage of level 1 items from 2003-04 to 2005-06. However, three of these teachers, Evelyn, Nancy, and Edward, still used over 90% level 1 items in 2005-06. Teachers overwhelmingly assessed students at depth of knowledge level 1. In fact, five of the nine teachers included over 90% of their items as level 1 on their assessments during both school years. Two of those five teachers (Evelyn and Edward) submitted tests from their advanced mathematics classes. Even the presence of a state test, one that includes items at levels 2 and 3, did not entice teachers to assess students at these higher depths of knowledge.

A few teachers used significant percentages of level 2 items or showed moderate increases in the depth of knowledge level of their items from 2003-04 to 2005-06. Sam

was the only teacher to have percentages similar to the OAT released items in both 2003-04 (61.9% level 1 and 35.1% level 2) and 2005-06 (58.2% level 1 and 38.5% level 2). Wanda and Linda showed some moderate decrease in level 1 items, accompanied by a similar increase in level 2 items, between 2003-04 and 2005-06. Wanda's assessments changed from 90% level 1 and 9.6% level 2 in 2003-04 to 83.9% level 1 and 16.1% level 2 in 2005-06. Linda's assessments changed from 87.3% level 1 and 11.3% level 2 in 2003-04 to 79.5% level 1 and 19.7% level 2 in 2005-06. Two of these three teachers, Sam and Linda, used *Connected Mathematics Project* (CMP) curriculum materials with their eighth-grade students and relied on these materials for their test items.

Data Standard

Data items from the teachers presented a different pattern than items assessing the other four standards. Teachers' items assessed the Data Standard with a higher depth of knowledge than other standards, with a larger proportion of level 2 items. In 2005-06, the ratio of level 2 to level 1 items in Data was close to 1:2. This was the only standard in either year in which teachers used this large of a proportion of its items above level 1. All other standards had a much lower ratio of level 2 to level 1 items. The ratio of data items in teachers' assessment items was consistent with data items from the OAT. Both the OAT items and teachers' items had a higher percentage of level 2 items in Data than any other standard.

Reliance on Curriculum Materials

Teachers relied heavily on textbook materials for selecting their assessment items. Seven of the nine teachers in the study reported that they relied on their curriculum materials to develop their tests. Interestingly, these teachers used a significant percentage

of items that assessed content considered below the eighth-grade (over one-third of their items). In addition, a large percentage of the classroom assessment items were level 1 (approximately 86%). Curriculum materials seemed to serve as poor sources for teachers trying to assess students at higher depth of knowledge levels.

Only two teachers deviated from this practice; they used *Connected Mathematics Project (CMP)*, an NSF-funded program. Sam and Linda used the lowest percentages of level 1 items in 2005-06. Both teachers reported that they relied on CMP items for their test items. Test items from Sam and Linda, and consequently from CMP, assessed students at higher depth of knowledge levels than other curriculum materials.

The two teachers that did not rely on their curriculum materials demonstrated significant changes in assessments between 2003-04 and 2005-06. Both teachers indicated that their items better addressed the indicators and OAT in 2005-06; however, the data from classroom assessments revealed otherwise. Frank stated that his items were more focused on the indicators and more like standardized test items in 2005-06. However, his items were less representative of the indicators and assessed students at lower depths of knowledge. The percentage of items assessing content below the eighth-grade level increased from 3% in 2003-04 to 17.1% in 2005-06. The percentage of level 1 items increased from 64.4% to 82.9%. Both of these findings illustrate how Frank's assessments changed in the opposite direction from his prediction.

Edward's items also changed significantly from 2003-04 to 2005-06. Many of these changes were related to a change in curriculum materials and a shift from an algebra course in 2003-04 to an integrated mathematics course in 2005-06. Edward reported that his items better reflected the eighth-grade indicators and OAT. However,

the analysis revealed that Edward's tests changed in ways that were less aligned with the indicators and OAT. Edward increased the percentage of items that assessed content below eighth grade from 16% in 2003-04 to 35.6% in 2005-06. Despite teaching the equivalent of a ninth-grade course, over one-third of the items assessed content below eighth grade in 2005-06. In terms of depth, over 90% of Edward's items were level 1 in both 2003-04 and 2005-06. In terms of mathematics content and depth of knowledge, Edward's items did not reflect alignment with the indicators and OAT as he predicted.

Discussion

This section provides a discussion about the key findings of the study, including connections among findings from the assessment analyses, the interview analyses, and previous research.

Low Depth of Knowledge and Below Eighth-Grade Content

One of the most disturbing findings of the study was that a large percentage of classroom assessment items were classified at low depths of knowledge. Unfortunately, this finding was consistent with earlier findings on the effects of state tests (Shepard, 2000; Smith, 1991). Teachers in this study included large percentages of items that reduced mathematics tasks to their simplest components and sought simplistic answers (Smith, 1991). This practice was true for teachers' 2003-04 items, and did not change as a result of the introduction of the state test in March 2005. Teachers continued to assess students at low levels despite a state test that assessed students at greater depths of knowledge like conceptual understanding and strategic thinking. State tests focusing on greater depths of knowledge did not seem to encourage teachers to assess students at greater depths of knowledge in mathematics.

A second disturbing finding was the relatively large percentage of teachers' items that assessed content below the eighth-grade level. This percentage was large (28.5%) in the year prior to the state test and actually increased to an average of almost one-third of teachers' items in 2005-06. Teachers seemed unable to interpret the eighth-grade indicators, and the state tests items that assessed them, in order to write larger percentages of classroom assessment items that addressed eighth-grade content.

Both of these findings, along with teachers' responses to interviews, point to teachers' heavy reliance on curriculum materials. While textbook companies and authors claim to align with Ohio's ACS, many simply are not aligned as closely as they need to be. Many items on teachers' classroom assessments were developed by the authors and publishers of their curriculum materials. In fact many of the classroom assessments were pre-made by the publisher of the materials. These pre-made tests did not assess students at a variety of depths of knowledge and assessed a significant percentage of content below the eighth grade.

The fact that most teachers relied on curriculum materials to write test items is consistent with findings by Brown, Cooney, and Jones (1990). Teachers in this study seemed more than willing to rely on external authority (textbook authors and publishers) to determine the way they assessed students. The low depth of knowledge of classroom assessment items is also consistent with previous research. Cooney, Badger, and Wilson (1993) found that teachers were unable to write items that provided insight into student thinking. In addition, Black and Wiliam (1998) found that state tests were often poor models for teachers. These findings appear consistent in terms of the released OAT items

in that 61.8% of the items were at depth of knowledge level 1. However, the released OAT items had lower percentages than teachers' items in this regard.

What is clear from this study is that the curriculum materials provided poor assessment models for teachers, both in terms of mathematics content and depth of knowledge. All of the teachers except for Frank and Edward relied on their curriculum materials for their test items, and these items assessed students at a very low depth of knowledge. Unfortunately, the analysis of teachers' interviews seemed to indicate that they were unaware of the mismatch between curriculum assessments and ACS. In fact, most teachers discussed how they challenged students with the assessment items from their curriculum materials. They simply seemed unable to address the issue of depth of knowledge. For example, items they considered challenging were often rigorous or multi-step, but they still assessed students' ability to memorize facts or carry out routine procedures. Their classroom assessment items simply were not challenging in terms of the depth of knowledge.

Further support that curriculum materials influenced teachers' assessment decisions were the two teachers who used CMP materials. These teachers assessed students at higher depth of knowledge levels because the CMP materials included assessments focusing on higher depths of knowledge. Sam and Linda both relied on CMP materials for their assessment items, and these materials included a larger percentage of items that assessed students' conceptual and strategic thinking than other curriculum materials.

Regardless of whether teachers relied heavily on curriculum materials or not, they must be able to recognize which assessment items are assessing appropriate eighth-grade

content and which are not. The teachers in this study seemed unable to interpret Ohio's indicators and align their assessments to these indicators. Six of the nine teachers specifically mentioned the Ohio ACS as influencing the mathematics content that they assess. However, these six teachers still assessed content below the eighth grade with significant percentages of items. These significant percentages may be related to teachers' oversimplifying the mathematics content of the indicators. For example, the eighth-grade indicators in the Number Standard recommend that students use ratios or rational numbers to solve problems. An item that asked students to compute a tip at a restaurant aligned with this indicator. However, items that required students to simplify number sentences with rational numbers aligned with indicators at the fifth- and sixth-grade level. Indicators at these grades focused on students' ability to compute fluently with fractions.

Regardless of the sources on which teachers rely for assessments, they should recognize that a significant percentage of their assessment items only assess low depths of knowledge. Assessments reflect what teachers believe is important (Cooney, Badger, & Wilson, 1993; NCTM, 2000). The large percentage of level 1 items seemed related to teachers beliefs about mathematics. By using a large percentage of level 1 items, teachers seemed to believe that doing mathematics entailed carrying out standard, well-known procedures and recalling factual information. This view is consistent with the traditional view of mathematics described by Brown, Cooney, and Jones (1990). The tests provided with curriculum materials, in addition to being easier to use, supported and validated these belief about mathematics. For example, during his interview, Henry

specifically mentioned using multiple-choice and fill-in-the-blank items because they were easier to grade.

Assessing More Content.

Previous research on state tests found that teachers tended to narrow the content that they taught, emphasizing only content that was included in the state test (Abrams, Pedulla, & Madaus, 2003; Firestone, Mayrowetz, & Fairman, 1998; Taylor, Shepard, Kinner, & Rosenthal, 2002). The findings of this study were not consistent with these findings. The nine teachers in this study actually increased the content that they assessed from 2003-04 to 2005-06 in response to the state test. They averaged assessing two additional content categories after the introduction of the state test. The increase in categories came largely from teachers' overemphasis on algebra and number content during 2003-04. Items in 2005-06 assessed more of the content categories, although large percentages of items continued to overemphasize algebra and number content. In 2005-06, teachers did not reduce the number of algebra or number items, rather they added new content like geometry, data analysis, and probability found on the OAT. In some cases, teachers simply added some released OAT items to their set of classroom assessments. While additional categories were assessed, teachers failed to assess some categories with a significant percentage of items.

Short-Cycle Assessments.

Frank's tests went through significant changes between the 2003-04 and 2005-06 school years. The 2005-06 tests were developed by the curriculum leader in Frank's district and were given to all students in regular eighth-grade mathematics classes. These short-cycle assessments were developed specifically to prepare students to be successful

on the OAT. The changes in Frank's assessment items affected not only Frank's classroom assessments, but all assessments used in eighth-grade mathematics classes in his district. Frank felt that the changes in his assessments were positive. He thought that the 2005-06 items better addressed the eighth-grade indicators and looked more like standardized tests. The findings of this study, however, revealed these changes resulted in a larger percentage of items that assessed content below eighth grade and at level 1 depth of knowledge. Frank, as well as other mathematics teachers in the study, used assessments that actually aligned less with state standards and assessments than in previous years. Like other teachers in this study, the curriculum leader in Frank's district also seemed unable to interpret accurately the recommended mathematics content for eighth-grade students..

Prior to the state test, Frank's items were relatively challenging in terms of depth of knowledge, with almost 15% of his items as level 3. The curriculum leader, however, in attempting to align with state tests and better prepare his teachers for the state test, included a larger percentage of level 1 items. This practice is consistent with what Smith (1991) described as reducing mathematics to its most basic set of individual skills and repeatedly drilling these isolated skills for mastery. This practice also perpetuates the belief that doing mathematics is only memorizing facts and performing routine procedures.

Implications

This section provides a summary of implications that can be drawn from the findings of study for practicing teachers, administrators, professional development providers, and researchers.

Implications for Practicing Teachers and Professional Development Providers

The teachers in this study clearly needed help in choosing and writing assessment items that (1) reflected the appropriate content at their grade level and (2) assessed students' knowledge beyond an ability to recall information and perform traditional procedures. The teachers needed assistance in crafting items that reflect deeper levels of knowledge like conceptual understanding and strategic thinking. Not only did they not seem aware of the content and depth of knowledge items in relation to the ACS and OAT, but they were unable to modify or find items that assessed eighth-grade content at appropriate depths of knowledge.

Professional development for mathematics teachers needs to assist teachers in analyzing the mathematics content and depths of knowledge of assessment tasks. Although Webb's (1999) framework provides a good example, others developed by Porter (2002) and Smith and Stein (1998) also exist. Teachers need experience at recognizing the qualities of an item that make it conceptual instead of procedural. In addition to identifying depth of knowledge levels for items, professional development should focus on giving teachers experience in writing or amending assessment items to focus on various depths of knowledge. Teachers must come to understand the value of including items at various depths of knowledge in assessing students in mathematics. They must also learn how to align these items to the indicators that serve as learning targets.

Through this professional development opportunity, hopefully teachers will develop more sophisticated beliefs about what it means to learn and do mathematics. This study and other literature (Cooney, Badger, & Wilson, 1993; NCTM, 2000) confirm that

assessments reflect what teachers value. Without taking time through professional development opportunities to explore and develop these beliefs, teachers will continue to only assess students' ability to recall facts and carry out routine procedures.

Implications for School Administrators

School administrators in Ohio, as well as administrators everywhere, are concerned about student achievement. Test results are often reported in newspapers and used to compare school buildings and districts. Schools are judged largely on student performance. Administrators need to be cautious, however, in the ways that they support teachers to improve student achievement. Assuming the findings of this study represent the practices of other teachers across the country, administrators must consider the strategies they use to support teachers. Simply put, the teachers in this study need to develop a more comprehensive view of assessment in mathematics.

Using Frank's administrator as an example, school administrators must ensure that they avoid doing harm. They must make sure that mathematics teachers expand, rather than narrow, their classroom assessment practices. This practice may be at odds with administrators' views on raising student achievement. They may be placing pressure on teachers to focus on preparing students for the test in ways that encourage lower depths of knowledge assessment items. Many administrators, like teachers, will most likely need to broaden their view of what mathematics assessments should look like and what roles it should take. Also like teachers, administrators need exposure to depth of knowledge levels and the importance of assessing students beyond memorization and performing procedures. This task may be more important, and difficult, than the recommendations for professional development for teachers. First, administrators are

likely to have even less of a background in mathematics than the teachers they need to support in this area. Second, principals often control the means to which teachers can pursue professional development. In order for teachers to get the professional development they need, building administrators will need to be educated first.

Implications for future research

More research analyzing the classroom assessment practices of mathematics teachers is needed. This study raised several important questions for future research. First, would analysis of all classroom assessments, rather than just tests, yield different results? Perhaps teachers use higher depth-of-knowledge items as part of other classroom assessments, such as homework or project assignments. A subsequent study could focus on all assessments used by teachers, ranging from large-scale projects to informal questions asked by teachers during lessons. More time would be spent with each teacher to understand the more complete classroom assessment picture. Through observations and interviews, one would need to gauge the importance placed on projects, portfolios, or other forms of assessments that are more integrated with instruction. Some tasks may be weighted more heavily than others in the coding process. Tasks could be coded with respect to content and depth of knowledge categories similar to this study.

Second, how do curriculum materials and classroom assessment practices relate? This question is particularly interesting considering the differences in assessments used by teachers using the NSF-funded *Connected Mathematics Project* materials and teachers using traditional textbooks. This study would entail an item analysis of the assessment items that are included with the most popular curriculum materials and those in CMP and other NSF-funded programs. Items could be examined in terms of mathematics content

and depth of knowledge and compared. In addition, are teachers as likely to use the assessments provided with different types of curriculum materials? If CMP materials include items with higher depths of knowledge, are teachers as likely to use them as they are items included at level 1?

Third, to what extent are districts moving towards common, short-cycle assessments and what strategies are districts using to align classroom and short-cycle assessments with state assessments? Administrators in Frank's district felt their short cycle assessments were better assessing the eighth-grade indicators and more aligned with the OAT. This study found the opposite to be true. Short cycle assessments from districts across Ohio could be collected and coded to examine this practice and determine if some short-cycle assessments were better aligned with the OAT. In addition, many teachers and school administrators would be interested to see if relationships between the content and depth of knowledge of short cycle assessments and student performance on state tests exist. A longitudinal study focused on these relationships would provide interesting results for policies with regard to district and classroom assessments. Each iteration of the short cycle assessments could be compared to respective yearly student achievement data.

Fourth, would the results of a study with different or larger samples of teachers produce different results? Larger samples of teachers would hopefully reveal a wider range of classroom assessment practices. The present study was limited because it only compared two years of assessment data (not every teacher could participate because they had not kept their assessments from two years prior to the study). Future studies might focus only on assessment practices with a larger group of teachers over longer periods of

time. A larger longitudinal study would enhance the generalizability of the results across a wider range of teachers and contexts.

Closing Remarks

This study provided an introductory view into the classroom assessment practices of nine middle school mathematics teachers. The teachers in this study, and the materials they used to assess students, generally did a poor job of preparing students for state assessments. Across the board, teachers and curriculum materials tended to assess students at below grade level and at low depths of knowledge. At least within the first two years, the introduction of the eighth-grade achievement test in Ohio did not entice teachers to change or expand their assessment practices with respect to the depth of knowledge and grade level mathematics content. Teachers' beliefs about mathematics, and their related assessment practices, continued to hinge on memorization of facts and performing routine procedures. This study has raised many questions, and many are left to be pursued. At a time of increased accountability, other studies that investigate the effects of state-mandated testing on classroom practices are needed.

REFERENCES

Abrams, L. M., Pedulla, J. J., Madaus, G. F. (2003). Views from the classroom: Teachers' opinions of statewide testing programs. *Theory into Practice, 42,* 18-29.

Ball, D.L. (1988). Unlearning to teach mathematics. *For the Learning of Mathematics, 8,* 40-48.

Behuniak, P. (2003). Education assessment in an era of accountability. In J. E. Wall & G. R. Walz (Eds.), *Measuring up: Assessment issues for teachers, counselors, and administrators* (pp. 335-47)

Black & Wiliam (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*, 139-147.

Bracey, G. W. (1987). Measurement-driven instruction: Catchy phrase, dangerous practice. *Phi Delta Kappan, 68*, 683-86.

Brown, S. I., Cooney, T. J., & Jones, D. (1990). Mathematics teacher education. In W. R. Houston (Ed.), *Handbook of research on teacher education* (pp. 639-656). New York: MacMillan.

Cohen, D. K. & Ball, D. L. (1990). Policy and practice: An overview. *Educational Evaluation and Policy Analysis, 12*, 233-39.

Cooney, T. J., Badger, E., Wilson, M.R. (1993). Assessment, understanding mathematics, and distinguishing visions from mirages. In N. Webb (Ed.), *Assessment in the mathematics classroom: 1993 yearbook* (pp. 239-247). Reston, VA: National Council of Teachers of Mathematics.

Corbett, H. D. & Wilson, B. L. (1991). Two state minimum competency testing programs and their effects on curriculum and instruction. Philadelphia, PA: Research for Better Schools, Inc.

Darling-Hammond, L. (2004). Standards, accountability, and school reform. *Teachers College Record, 106*, 1047-1085.

DeBoer, G. (2004). High-quality assessment items on the horizon. *2061 Connections*. Retrieved on January 5, 2006, from http://www.project2061.org/publications/2061Connections/ 2004/2004-05c.htm

Dossey, J. A. (1992). The nature of mathematics: Its role and its influence. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 39-48). New York: Maxwell Macmillan International.

Firestone, W. A., Mayrowetz, D., & Fairman, J. (1998). Performance-based assessments and instructional change: The effects of testing in Maine and Maryland. *Educational Evaluation and Policy Analysis, 20,* 95-113.

Fremer, & Wall. (2003). Why use tests and assessments? In J. E. Wall & G. R. Walz (Eds.), *Measuring up: Assessment issues for teachers, counselors, and administrators,* 3-19:

Gallagher, C. J. (2003). Reconciling a tradition of testing with a new paradigm. *Educational Psychology Review, 15*, 83-99.

Glassnapp, D. R., Poggio, J. P., & Miller, D. M. (1991). Impact of a "low stakes" state minimum competency testing program on policy, attitudes, and achievement. In R. E. Stake (Ed.), *Advances in program evaluation: Vol. 1. Effects of mandated assessment on teaching* (pp. 101-140). Greenwich, CT: JAI Press Ltd.

Grant, S. G. (2000). Teachers and tests: Exploring teachers' perceptions of changes in the New York state testing program. *Education Policy Analysis Archives, 8(14).*

Hanson, F. A. (1993). Testing, testing: Social consequences of the examined life. Berkeley, CA: University of California Press

Hoy, A. W., Hoy, W. K., (2003). Instructional leadership: A learning-centered guide. Boston, MA: Pearson.

Jaberg, P., Lubinski, C., Aeschleman, S. (2004). Developing a support system for teacher change in mathematics education: The principal's role. In R. N. Rubenstein & G. W. Bright (Eds.), *Perspectives on the teaching of mathematics: 2004 yearbook* (pp. 229-38). Reston, VA: National Council of Teachers of Mathematics.

Joram, E. & Gabriele, A. J. (1998). Pre-service teachers' prior beliefs: Transforming obstacles into opportunities. *Teaching and Teacher Education, 14,*175-191.

Kagan, D. M. (1992). Implications of research on teacher belief. *Educational Psychologist, 27*, 65-90.

Koretz, D., Stecher, B., Klein, S., McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice, 13(3),* 5-16.

Kuppermintz, H., Shepard, L. A., & Linn, R. (2001, April). *Teacher effects as a measure of teacher effectiveness: Construct validity considerations in TVAAS (Tennessee*

Value Added Assessment System). Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA.

Lerman, S. (1997, July). The psychology of mathematics teachers' learning: In search of theory. Proceedings of the Conference of the International Group for the Psychology of Mathematics Education, Lahti, Finland.

Lonergan. (2003). A Guide to the Testing and Accountability Requirements of No Child Left Behind. For Parents, about Parents. Washington, DC: Office of Educational Research and Improvement.

Louis, K., Kruse, S., & Marks, H. (1996). Schoolwide professional community. In Newmann, F., & Associates (Eds.), *Authentic achievement: restructure schools for intellectual quality* (pp.161-178). San Francisco, CA: Jossey-Bass Publishers.

Ma, L. (1999). *Knowing and teaching elementary mathematics*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mewborn, D. S. (2002, April). Examining mathematics teachers' beliefs through multiple lenses. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

McBee, R. H. (2002). When it comes to testing, why not make lemonade? *Educational Forum, 66,* 238-246

McMillan, J.H., Myran, S., & Workman, D. (1999, April). The impact of mandated statewide testing on teachers' classroom assessment and instructional practices. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.

National Center for Educational Statistics (2002). Navigating Resources for Rural Schools, retrieved July 20, 2006, from http://nces.ed.gov/surveys/RuralEd/definitions.asp#Locale

National Council of Teachers of Mathematics (1989). *Curriculum and evaluation standards for school mathematics*. NCTM: Reston, VA

National Council of Teachers of Mathematics (1995). *Assessment standards for school mathematics*. NCTM: Reston, VA

National Council of Teachers of Mathematics (2000). *Principals and standards for school mathematics*. NCTM: Reston, VA

National Middle School Association (2003). *This we believe: Successful schools for young adolescents*. Columbus, OH: National Middle School Association.

Ohio Department of Education (1990). High School Proficiency Testing: Fact Sheets, Mathematics. Retrieved on January 5, 2006, from http://www.ode.state.oh.us/proficiency/sample_tests/ninth/9fsmath.pdf

Ohio Department of Education (2002). *Academic content standards: K-12 mathematics.* Columbus, OH: Author

Ohio Department of Education (2004). Ohio Grade 8 Mathematics Achievement Test Blueprint. Office of Assessment, and Office of Curriculum and Instruction Mathematics Team.

Ohio Department of Education (2006a). Information guide to Ohio proficiency tests for grade 6: Mathematics. Retrieved January 5, 2006, from http://www.ode.state.oh.us/proficiency/sample_tests/sixth/6rmath.pdf

Ohio Department of Education (2006b). Information guide to Ohio proficiency tests for grade 4: Mathematics. Retrieved January 5, 2006, from http://www.ode.state.oh.us/proficiency/sample_tests/fourth/4rmath.pdf

Olson, L. (2004). Testing. *Education Week*, *23(23),*17.

Paris, G. P., Lawton, T. A., Turner, J. C., Roth, J. L. (1991). A developmental perspective on standardized achievement testing. *Educational Researcher, 20(5),* 12-20.

Pajares, M. F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research, 62,* 307-322.

Popham, W. J. (1987). The merits of measurement-driven instruction. *Phi Delta Kappan, 68,* 679-682.

Popham, W. J. (2003). Trouble with testing. *American School Board Journal, 190(2),* 14-17.

Porter, A. C. (2002). Measuring the Content of Instruction: Uses in Research and Practice. *Educational Researcher, 31(7),* 3-14.

Rose, L. C., & Gallup, A.M. (2001). The 33rd annual Phi Delta Kappan/Gallup poll of the public's attitude toward the public schools. *Phi Delta Kappan*, *83*, 41-58.

Rousseau, C. K. (2004). Shared beliefs, conflict, and a retreat from reform: The story of a professional community of high school mathematics teachers. *Teaching and Teacher Education (20)*, 783-796.

Schorr, R. Y., Firestone, W. A., & Monfils, L. (2003). State testing and mathematics teaching in New Jersey: The effects of a test without other supports. *Journal for Research in Mathematics Education, 34,* 373-405.

Schulman, L. (1996). New assessment practices in mathematics. *Journal of Education, 178*, 61-71.

Shepard, L. (2000). The role of classroom assessment in teaching and learning. Los Angeles, CA: California University, Center for Research on Evaluation, Standards, and Student Testing.

Shulman, L. (1986). Those who understand: knowledge growth in teaching. *Educational Researcher, 10(1)*, 4-14.

Smith, M.L. (1991). Put to the test: The effects of external testing on teachers. *Educational Researcher, 20(5),* 8-11

Smith, M. S. & Stein, M. K. (1998). Selecting and creating mathematical tasks: From research to practice. *Mathematics Teaching in the Middle School, 3(5),* 344-350.

Stein , M. K. & Lane, S. (1996, April). Classrooms in which students successfully acquire mathematical proficiency: What are the critical features of teachers' instructional practice? Paper presented at the annual meeting of the American Educational Research Association, New York.

Stiggins, R. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan, 83,* 758-65.

Stiggins, R. (2004). New assessment beliefs for a new school mission. *Phi Delta Kappan*, 86, 22-27.

Stigler, J. & Hiebert, J. (1999). The teaching gap. New York, NY: The Free Press.

Taylor, G., Shepard, L., Kinner, F., Rosenthal, J. (2002). A survey of teachers' perspectives on high-stakes testing in colorado: What gets taught, what gets lost. Los Angeles, CA: California University, Center for the Study of Evaluation.

Taylor, P. (2004). Encouraging professional growth and mathematics reform through collegial interaction. In R. N. Rubenstein & G. W. Bright (Eds.), *Perspectives on the teaching of mathematics: 2004 yearbook* (pp. 219-28). Reston, VA: National Council of Teachers of Mathematics.

Thompson, A. G. (1992). Teachers' beliefs and conceptions: a synthesis of the research. In D. A. Grouws (Ed.), Handbook of research on mathematics teaching and learning (pp. 39-48). New York: Maxwell Macmillan International.

Urdan, T.C., & Paris, S.G. (1994). Teachers' perceptions of standardized achievement tests. *Educational Policy, 8*, 137-156

United States Department of Education (2005). Helping families, schools and communities understand and improve student achievement. Retrieved on November 6, 2005, from http://www.ed.gov/nclb/accountability/ayp/testingforresults.html

Walton, S. & Taylor, K. (1997). How did you know the answer was a boxcar? *Educational Leadership, 54(4),* 38–40.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states* (Research Monograph No. 18). National Institute for Science Education. Madison, WI: University of Wisconsin-Madison, National Institute for Science Education.

Wiliam, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for Learning: Impact on student achievement. *Assessment in Education: Principles, Policy and Practice, 11*, 49-65.

Wilson, S. (2003). *California dreaming: Reforming mathematics education*. New Haven, CT: Yale University Press.

APPENDICES

APPENDIX A
Locale codes for schools and districts
(NCES, 2002)

1. Large City - A central city of a Core Based Statistical Area (CBSA) or Consolidated Statistical Area (CSA), with the city having a population greater than or equal to 250,00.
2. Mid-size City - A central city of a CBSA or CSA, with the city having a population less than 250,000.
3. Urban Fringe of a Large City - Any incorporated place, Census Designated Place, or non-place territory within a CBSA or CSA of a Large City and defined as urban by the Census Bureau.
4. Urban Fringe of a Mid-size City - Any incorporated place, Census Designated Place, or non-place territory within a CBSA or CSA of a Mid-size City and defined as urban by the Census Bureau.
5. Large Town - An incorporated place or Census Designated Place with a population greater than or equal to 25,000 and located outside a CBSA or CSA.
6. Small Town - An incorporated place or Census Designated Place with a population less than 25,000 and greater than or equal to 2,500 and located outside a CBSA or CSA.
7. Rural, outside CBSA - Any incorporated place, Census designated place, or non-place territory not within a CBSA or CSA of a Large or Mid-size City and defined as rural by the Census Bureau.
8. Rural, inside CBSA - Any incorporated place, Census designated place, or non-place territory within a CBSA or CSA of a Large or Mid-size City and defined as rural by the Census Bureau.

APPENDIX B
Eighth Grade Indicators and Content Categories
(ODE, 2002)

Grade Eight

Number, Number Sense and Operations Standard

Number and Number Systems

1. Use scientific notation to express large numbers and small numbers between 0 and 1.
2. Recognize that natural numbers, whole numbers, integers, rational numbers and irrational numbers are subsets of the real number system.

Meaning of Operations

3. Apply order of operations to simplify expressions and perform computations involving integer exponents and radicals.
4. Explain and use the inverse and identity properties and use inverse relationships (addition/subtraction, multiplication/division, squaring/square roots) in problem solving situations.

Computation and Estimation

5. Determine when an estimate is sufficient and when an exact answer is needed in problem situations, and evaluate estimates in relation to actual answers; e.g., very close, less than, greater than.
6. Estimate, compute and solve problems involving rational numbers, including ratio, proportion and percent, and judge the reasonableness of solutions.
7. Find the square root of perfect squares, and approximate the square root of non-perfect squares as consecutive integers between which the root lies; e.g., $\sqrt{130}$ is between 11 and 12.
8. Add, subtract, multiply, divide and compare numbers written in scientific notation.

Measurement Standard

Measurement Units

1. Compare and order the relative size of common U.S. customary units and metric units; e.g., mile and kilometer, gallon and liter, pound and kilogram.
2. Use proportional relationships and formulas to convert units from one measurement system to another; e.g., degrees Fahrenheit to degrees Celsius.

Use Measurement Techniques and Tools

3. Use appropriate levels of precision when calculating with measurements.

4. Derive formulas for surface area and volume and justify them using geometric models and common materials. For example, find:

   a. the surface area of a cylinder as a function of its height and radius;
   b. that the volume of a pyramid (or cone) is one-third of the volume of a prism (or cylinder) with the same base area and height.

5. Determine surface area for pyramids by analyzing their parts.

6. Solve and determine the reasonableness of the results for problems involving rates and derived measurements, such as velocity and density, using formulas, models and graphs.

7. Apply proportional reasoning to solve problems involving indirect measurements or rates.

8. Find the sum of the interior and exterior angles of regular convex polygons with and without measuring the angles with a protractor.

9. Demonstrate understanding of the concepts of perimeter, circumference and area by using established formulas for triangles, quadrilaterals, and circles to determine the surface area and volume of prisms, pyramids, cylinders, spheres and cones. (Note: Only volume should be calculated for spheres and cones.)

10. Use conventional formulas to find the surface area and volume of prisms, pyramids and cylinders and the volume of spheres and cones to a specified level of precision.

Geometry and Spatial Sense Standard

Characteristics and Properties

1. Make and test conjectures about characteristics and properties (e.g., sides, angles, symmetry) of two-dimensional figures and three-dimensional objects.

2. Recognize the angles formed and the relationship between the angles when two lines intersect and when parallel lines are cut by a transversal.

3. Use proportions in several forms to solve problems involving similar figures (part-to-part, part-to-whole, corresponding sides between figures).

Spatial Relationships

4. Represent and analyze shapes using coordinate geometry; e.g., given three vertices and the type of quadrilateral, find the coordinates of the fourth vertex.

Transformations and Symmetry

5. Draw the results of translations, reflections, rotations and dilations of objects in the coordinate plane, and determine properties that remain fixed; e.g., lengths of sides remain the same under translations.

Visualization and Geometric Models

6. Draw nets for a variety of prisms, pyramids, cylinders and cones.

Patterns, Functions and Algebra Standard

Use Patterns, Relations and Functions

1. Relate the various representations of a relationship; i.e., relate a table to graph, description and symbolic form.
2. Generalize patterns and sequences by describing how to find the *n*th term.
3. Identify functions as linear or nonlinear based on information given in a table, graph or equation.

Use Algebraic Representations

4. Extend the uses of variables to include covariants where y depends on x.
5. Use physical models to add and subtract monomials and polynomials, and to multiply a polynomial by a monomial.
6. Describe the relationship between the graph of a line and its equation, including being able to explain the meaning of slope as a constant rate of change and y-intercept in real-world problems.
7. Use symbolic algebra (equations and inequalities), graphs and tables to represent situations and solve problems.
8. Write, simplify and evaluate algebraic expressions (including formulas) to generalize situations and solve problems.
9. Solve linear equations and inequalities graphically, symbolically and using technology.
10. Solve 2 by 2 systems of linear equations graphically and by simple substitution.
11. Interpret the meaning of the solution of a 2 by 2 system of equations; i.e., point, line, no solution.

12. Solve simple quadratic equations graphically; e.g., $y = x^2 - 16$.

13. Compute and interpret slope, midpoint and distance given a set of ordered pairs.

Analyze Change

14. Differentiate and explain types of changes in mathematical relationships, such as linear vs. nonlinear, continuous vs. noncontinuous, direct variation vs. inverse variation.

15. Describe and compare how changes in an equation affects the related graphs; e.g., for a linear equation changing the coefficient of *x* affects the slope and changing the constant affects the intercepts.

16. *Use graphing calculators or computers to analyze change; e.g., interest compounded over time as a nonlinear growth pattern.*

Data Analysis and Probability Standard

Data Collection

1. Use, create and interpret scatterplots and other types of graphs as appropriate.

2. Evaluate different graphical representations of the same data to determine which is the most appropriate representation for an identified purpose; e.g., line graph for change over time, circle graph for part-to-whole comparison, scatterplot for relationship between two variants.

3. Differentiate between discrete and continuous data and appropriate ways to represent each.

Statistical Methods

4. Compare two sets of data using measures of center (mean, mode, median) and measures of spread (range, quartiles, interquartile range, percentiles).

5. Explain the mean's sensitivity to extremes and its use in comparison with the median and mode.

6. Make conjectures about possible relationship in a scatterplot and approximate line of best fit.

7. Identify different ways of selecting samples, such as survey response, random sample, representative sample and convenience sample.

8. Describe how the relative size of a sample compared to the target population affects the validity of predictions.

9. Construct convincing arguments based on analysis of data and interpretation of graphs.

Probability

10. Calculate the number of possible outcomes for a situation, recognizing and accounting for when items may occur more than once or when order is important.

11. Demonstrate an understanding that the probability of either of two disjoint events occurring can be found by adding the probabilities for each and that the probability of one independent event following another can be found by multiplying the probabilities.

APPENDIX C
2003-04 Assessment Item Data

Table C1

2003-04 Assessment Item Data: Number and Measurement

Content Categories	Summary Data			Individual Teacher Data				
	Indicators N=51	OAT N=76	Mean	Sam N=97	Wanda N=261	Frank N=101	Evelyn N=236	Nancy N=300
Number Below Eighth Grade	0.0%	0.0%	14.3%	3.1%	9.2%	1.0%	5.5%	26.7%
Number and Number Systems	3.9%	2.6%	3.2%	1.0%	1.5%	6.9%	4.7%	2.7%
Meaning of Operation	3.9%	3.9%	3.1%	0.0%	6.5%	6.9%	3.8%	3.7%
Computation and Estimation	7.8%	11.8%	3.4%	2.1%	3.1%	8.9%	2.5%	5.0%
Number Above Eighth Grade	0.0%	0.0%	0.2%	0.0%	0.0%	0.0%	2.1%	0.0%
Measurement Below Eighth Grade	0.0%	0.0%	4.7%	6.2%	13.8%	2.0%	1.7%	2.3%
Measurement Units	3.9%	1.3%	1.7%	0.0%	0.8%	2.0%	0.0%	0.0%
Use Measurement Techniques and Tools	15.7%	18.4%	4.4%	0.0%	4.6%	10.9%	1.3%	0.7%
Measurement Above Eighth Grade	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

Table C2

2003-04 Assessment Item Data: Geometry and Algebra

Content Categories	Summary Data			Individual Teachers				
	Indicators	OAT	Mean	Sam	Wanda	Frank	Evelyn	Nancy
Geometry Below Eighth Grade	0.0%	1.3%	3.9%	6.2%	8.4%	0.0%	3.8%	2.3%
Characteristics and Properties	5.9%	10.5%	2.4%	0.0%	5.0%	5.9%	0.0%	1.0%
Spatial Relationships	2.0%	2.6%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Transformations and Symmetry	2.0%	3.9%	2.0%	0.0%	2.3%	4.0%	0.0%	3.3%
Visualization and Geometric Models	2.0%	1.3%	0.2%	0.0%	0.4%	1.0%	0.0%	0.0%
Geometry Above Eighth Grade	0.0%	0.0%	0.4%	0.0%	1.5%	0.0%	0.0%	0.0%
Algebra Below Eighth Grade	0.0%	0.0%	3.5%	13.4%	1.9%	0.0%	2.1%	4.3%
Use Patterns, Relations, and Functions	5.9%	5.3%	2.7%	16.5%	0.0%	5.9%	0.0%	0.0%
Use Algebraic Representations	19.6%	21.1%	28.6%	30.9%	19.5%	27.7%	25.4%	26.0%
Analyze Change	5.9%	2.6%	0.7%	4.1%	0.0%	2.0%	0.0%	0.0%
Algebra Above Eighth Grade	0.0%	0.0%	12.1%	17.5%	0.8%	0.0%	45.8%	2.7%

Table C3

2003-04 Assessment Item Data: Data

Content Categories	Summary Data			Individual Teachers				
	Indicators	OAT	Mean	Sam	Wanda	Frank	Evelyn	Nancy
Data Below Eighth Grade	0.0%	1.3%	2.1%	0.0%	6.5%	0.0%	0.0%	7.0%
Data Collection	5.9%	7.9%	3.4%	0.0%	8.8%	4.0%	1.3%	7.7%
Statistical Methods	11.8%	6.6%	1.6%	0.0%	0.4%	5.0%	0.4%	2.3%
Probability	3.9%	5.3%	1.0%	0.0%	1.9%	5.0%	0.0%	0.7%
Data Above Eighth Grade	0.0%	0.0%	0.8%	0.0%	4.2%	1.0%	0.0%	1.7%

Table C4

2003-04 Assessment Item Data: Number and Measurement

Content Categories	Summary Data			Individual Teachers			
	Indicators N=51	OAT N=76	Mean	Helen N=215	Henry N=1044	Edward N=213	Linda N=110
Number Below Eighth Grade	0.0%	0.0%	14.3%	27.0%	25.7%	10.8%	20.0%
Number and Number Systems	3.9%	2.6%	3.2%	2.3%	0.7%	0.5%	8.2%
Meaning of Operation	3.9%	3.9%	3.1%	0.5%	4.1%	0.5%	1.8%
Computation and Estimation	7.8%	11.8%	3.4%	4.7%	1.6%	1.4%	0.9%
Number Above Eighth Grade	0.0%	0.0%	0.2%	0.0%	0.0%	0.0%	0.0%
Measurement Below Eighth Grade	0.0%	0.0%	4.7%	9.8%	6.4%	0.0%	0.0%
Measurement Units	3.9%	1.3%	1.7%	0.9%	0.0%	0.0%	11.8%
Use Measurement Techniques and Tools	15.7%	18.4%	4.4%	9.3%	1.2%	0.5%	10.9%
Measurement Above Eighth Grade	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

Table C5

2003-04 Assessment Item Data: Geometry and Algebra

Content Categories	Summary Data			Individual Teachers			
	Indicators	OAT	Mean	Helen	Henry	Edward	Linda
Geometry Below Eighth Grade	0.0%	1.3%	3.9%	8.4%	5.7%	0.0%	0.0%
Characteristics and Properties	5.9%	10.5%	2.4%	6.0%	3.8%	0.0%	0.0%
Spatial Relationships	2.0%	2.6%	0.0%	0.0%	0.0%	0.0%	0.0%
Transformations and Symmetry	2.0%	3.9%	2.0%	7.4%	0.6%	0.0%	0.0%
Visualization and Geometric Models	2.0%	1.3%	0.2%	0.0%	0.0%	0.0%	0.0%
Geometry Above Eighth Grade	0.0%	0.0%	0.4%	2.3%	0.0%	0.0%	0.0%
Algebra Below Eighth Grade	0.0%	0.0%	3.5%	1.9%	2.9%	4.7%	0.0%
Use Patterns, Relations, and Functions	5.9%	5.3%	2.7%	0.5%	0.0%	0.0%	1.8%
Use Algebraic Representations	19.6%	21.1%	28.6%	9.8%	35.2%	49.8%	32.7%
Analyze Change	5.9%	2.6%	0.7%	0.0%	0.0%	0.0%	0.0%
Algebra Above Eighth Grade	0.0%	0.0%	12.1%	0.0%	9.8%	31.5%	0.9%

Table C6

2003-04 Assessment Item Data: Data

Content Categories	Summary Data		Individual Teachers				
	Indicators	OAT	Mean	Helen	Henry	Edward	Linda
Data Below Eighth Grade	0.0%	1.3%	2.1%	1.9%	0.8%	0.5%	2.7%
Data Collection	5.9%	7.9%	3.4%	5.1%	1.2%	0.0%	2.7%
Statistical Methods	11.8%	6.6%	1.6%	0.5%	0.3%	0.0%	5.5%
Probability	3.9%	5.3%	1.0%	1.9%	0.0%	0.0%	0.0%
Data Above Eighth Grade	0.0%	0.0%	0.8%	0.0%	0.0%	0.0%	0.0%

Table C7

2003-04 Assessment Item Data: Mathematics Content by Standard

		Standards				
Teachers	N	Number	Measurement	Geometry	Algebra	Data
Sam	97	6.2%	6.2%	6.2%	81.4%	0.0%
Wanda	261	20.3%	19.2%	17.6%	22.2%	21.5%
Frank	101	23.8%	14.9%	10.9%	35.6%	14.9%
Evelyn	236	18.6%	3.0%	3.8%	73.3%	1.7%
Nancy	300	38.0%	3.0%	6.7%	33.0%	19.3%
Helen	215	34.4%	20.0%	24.2%	12.1%	9.3%
Henry	1044	32.1%	7.7%	10.2%	47.8%	2.3%
Edward	213	13.1%	0.5%	0.0%	85.9%	0.5%
Linda	110	30.9%	22.7%	0.0%	35.5%	10.9%
Mean		24.2%	10.8%	8.8%	47.4%	8.9%
OAT	76	18.4%	19.7%	18.4%	26.3%	21.1%
Indicators	51	15.7%	19.6%	11.8%	31.4%	21.6%

Table C8

2003-04 Assessment Item Data: Below Grade Level Items

		Standards					
Teachers	N	Number	Measure	Geometry	Algebra	Data	Totals
Sam	97	3.1%	6.2%	6.2%	13.4%	0.0%	28.9%
Wanda	261	9.2%	13.8%	8.4%	1.9%	6.5%	39.8%
Frank	101	1.0%	2.0%	0.0%	0.0%	0.0%	3.0%
Evelyn	236	5.5%	1.7%	3.8%	2.1%	0.0%	13.1%
Nancy	300	26.7%	2.3%	2.3%	4.3%	7.0%	42.7%
Helen	215	27.0%	9.8%	8.4%	1.9%	1.9%	48.8%
Henry	1044	25.7%	6.4%	5.7%	2.9%	0.8%	41.5%
Edward	213	10.8%	0.0%	0.0%	4.7%	0.5%	16.0%
Linda	110	20.0%	0.0%	0.0%	0.0%	2.7%	22.7%
Mean		14.3%	4.7%	3.9%	3.5%	2.1%	28.5%

Table C9

2003-04 Assessment Item Data: Items Above the Eighth Grade

		Standards					
Teachers	N	Number	Measure	Geometry	Algebra	Data	Totals
Sam	97	0.0%	0.0%	0.0%	17.5%	0.0%	17.5%
Wanda	261	0.0%	0.0%	1.5%	0.8%	4.2%	6.5%
Frank	101	0.0%	0.0%	0.0%	0.0%	1.0%	1.0%
Evelyn	236	2.1%	0.0%	0.0%	45.8%	0.0%	47.9%
Nancy	300	0.0%	0.0%	0.0%	2.7%	1.7%	4.3%
Helen	215	0.0%	0.0%	2.3%	0.0%	0.0%	2.3%
Henry	1044	0.0%	0.0%	0.0%	9.8%	0.0%	9.8%
Edward	213	0.0%	0.0%	0.0%	31.5%	0.0%	31.5%
Linda	110	0.0%	0.0%	0.0%	0.9%	0.0%	0.9%
Mean		0.2%	0.0%	0.4%	12.1%	0.8%	13.5%

Table C10

2003-04 Assessment Item Data: Depth of Knowledge

		Depth of Knowledge		
Teachers	N	Level 1	Level 2	Level 3
Sam	97	61.9%	35.1%	3.1%
Wanda	261	90.0%	9.6%	0.4%
Frank	101	64.4%	20.8%	14.9%
Evelyn	236	97.9%	2.1%	0.0%
Nancy	300	93.0%	7.0%	0.0%
Helen	215	91.2%	8.8%	0.0%
Henry	1044	98.3%	1.7%	0.0%
Edward	213	95.3%	4.7%	0.0%
Linda	110	87.3%	11.8%	0.9%
Mean		86.6%	11.3%	2.1%
OAT	76	61.8%	35.5%	2.6%

APPENDIX D
2005-06 Assessment Item Data

Table D1

2005-06 Assessment Item Data: Number and Measurement

Content Categories	Summary Data			Individual Teacher Data				
	Indicators	OAT	Mean	Sam	Wanda	Frank	Evelyn	Nancy
	N=51	N=76		N=91	N=223	N=146	N=393	N=346
Number Below Eighth Grade	0.0%	0.0%	14.7%	4.4%	11.7%	5.5%	5.3%	24.0%
Number and Number Systems	3.9%	2.6%	2.4%	0.0%	0.4%	2.7%	1.0%	2.9%
Meaning of Operation	3.9%	3.9%	1.8%	0.0%	1.8%	2.7%	0.8%	4.0%
Computation and Estimation	7.8%	11.8%	4.4%	3.3%	4.0%	7.5%	1.3%	7.2%
Number Above Eighth Grade	0.0%	0.0%	0.4%	0.0%	0.0%	0.0%	0.5%	0.0%
Measurement Below Eighth Grade	0.0%	0.0%	5.1%	3.3%	7.6%	4.1%	0.3%	2.3%
Measurement Units	3.9%	1.3%	0.8%	0.0%	0.9%	2.1%	0.0%	0.0%
Use Measurement Techniques and Tools	15.7%	18.4%	3.4%	0.0%	4.9%	4.8%	1.8%	1.7%
Measurement Above Eighth Grade	0.0%	0.0%	0.1%	0.0%	0.0%	0.7%	0.0%	0.0%

2005-06 Assessment Item Data: Geometry and Algebra

Content Categories	Summary Data			Individual Teacher Data				
	Indicators	OAT	Mean	Sam	Wanda	Frank	Evelyn	Nancy
Geometry Below Eighth Grade	0.0%	1.3%	7.8%	6.6%	11.2%	3.4%	3.1%	2.3%
Characteristics and Properties	5.9%	10.5%	3.3%	0.0%	3.6%	3.4%	0.0%	1.4%
Spatial Relationships	2.0%	2.6%	0.1%	0.0%	0.0%	0.7%	0.0%	0.0%
Transformations and Symmetry	2.0%	3.9%	1.6%	0.0%	0.0%	0.7%	1.3%	3.8%
Visualization and Geometric Models	2.0%	1.3%	0.3%	0.0%	0.0%	0.7%	0.0%	0.3%
Geometry Above Eighth Grade	0.0%	0.0%	0.4%	0.0%	0.0%	0.0%	0.0%	0.0%
Algebra Below Eighth Grade	0.0%	0.0%	2.0%	2.2%	1.3%	2.7%	3.6%	3.8%
Use Patterns, Relations, and Functions	5.9%	5.3%	2.5%	14.3%	0.9%	2.1%	2.3%	0.6%
Use Algebraic Representations	19.6%	21.1%	25.3%	23.1%	29.1%	39.0%	27.5%	25.1%
Analyze Change	5.9%	2.6%	0.9%	5.5%	0.0%	2.7%	0.0%	0.3%
Algebra Above Eighth Grade	0.0%	0.0%	12.3%	18.7%	8.5%	3.4%	42.2%	2.3%

Table D3

2005-06 Assessment Item Data: Data

Content Categories	Summary Data			Individual Teacher Data				
	Indicators	OAT	Mean	Sam	Wanda	Frank	Evelyn	Nancy
Data Below Eighth Grade	0.0%	1.3%	3.0%	4.4%	4.9%	1.4%	1.0%	6.1%
Data Collection	5.9%	7.9%	3.6%	4.4%	4.0%	4.1%	5.9%	8.1%
Statistical Methods	11.8%	6.6%	2.4%	8.8%	2.2%	1.4%	0.5%	2.3%
Probability	3.9%	5.3%	0.9%	0.0%	0.9%	2.1%	1.0%	0.9%
Data Above Eighth Grade	0.0%	0.0%	1.1%	1.1%	2.2%	2.1%	1.5%	1.4%

2005-06 Assessment Item Data: Number and Measurement

Content Categories	Summary Data			Individual Teacher Data			
	Indicators N=51	OAT N=76	Mean	Helen N=933	Henry N=897	Edward N=177	Linda N=117
Number Below Eighth Grade	0.0%	0.0%	14.7%	36.0%	21.9%	13.0%	10.3%
Number and Number Systems	3.9%	2.6%	2.4%	1.1%	0.7%	3.4%	9.4%
Meaning of Operation	3.9%	3.9%	1.8%	1.1%	0.9%	3.4%	1.7%
Computation and Estimation	7.8%	11.8%	4.4%	4.1%	6.0%	4.0%	2.6%
Number Above Eighth Grade	0.0%	0.0%	0.4%	0.4%	2.1%	0.6%	0.0%
Measurement Below Eighth Grade	0.0%	0.0%	5.1%	8.3%	5.7%	5.1%	9.4%
Measurement Units	3.9%	1.3%	0.8%	0.2%	0.0%	0.0%	4.3%
Use Measurement Techniques and Tools	15.7%	18.4%	3.4%	4.3%	4.0%	5.1%	4.3%
Measurement Above Eighth Grade	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%

2005-06 Assessment Item Data: Geometry and Algebra

Content Categories	Summary Data			Individual Teacher Data			
	Indicators	OAT	Mean	Helen	Henry	Edward	Linda
Geometry Below Eighth Grade	0.0%	1.3%	7.8%	12.0%	9.4%	14.1%	7.7%
Characteristics and Properties	5.9%	10.5%	3.3%	2.4%	2.7%	11.9%	4.3%
Spatial Relationships	2.0%	2.6%	0.1%	0.0%	0.0%	0.0%	0.0%
Transformations and Symmetry	2.0%	3.9%	1.6%	2.1%	0.0%	0.6%	6.0%
Visualization and Geometric Models	2.0%	1.3%	0.3%	0.1%	0.0%	0.6%	0.9%
Geometry Above Eighth Grade	0.0%	0.0%	0.4%	0.6%	3.1%	0.0%	0.0%
Algebra Below Eighth Grade	0.0%	0.0%	2.0%	1.9%	2.2%	0.0%	0.0%
Use Patterns, Relations, and Functions	5.9%	5.3%	2.5%	0.1%	0.2%	0.0%	1.7%
Use Algebraic Representations	19.6%	21.1%	25.3%	12.2%	17.3%	24.3%	29.9%
Analyze Change	5.9%	2.6%	0.9%	0.0%	0.0%	0.0%	0.0%
Algebra Above Eighth Grade	0.0%	0.0%	12.3%	0.6%	22.4%	11.9%	0.9%

Table D6

2005-06 Assessment Item Data: Data

Content Categories	Summary Data			Individual Teacher Data			
	Indicators	OAT	Mean	Helen	Henry	Edward	Linda
Data Below Eighth Grade	0.0%	1.3%	3.0%	4.8%	0.8%	3.4%	0.0%
Data Collection	5.9%	7.9%	3.6%	2.9%	0.3%	1.1%	1.7%
Statistical Methods	11.8%	6.6%	2.4%	0.8%	0.4%	0.0%	5.1%
Probability	3.9%	5.3%	0.9%	2.3%	0.0%	0.6%	0.0%
Data Above Eighth Grade	0.0%	0.0%	1.1%	1.7%	0.0%	0.0%	0.0%

Table D7

2005-06 Assessment Item Data: Mathematics Content by Standard

		Standards				
Teachers	N	Number	Measurement	Geometry	Algebra	Data
Sam	91	7.7%	3.3%	6.6%	63.7%	18.7%
Wanda	223	17.9%	13.5%	14.8%	39.9%	14.3%
Frank	146	18.5%	11.6%	8.9%	50.0%	11.0%
Evelyn	393	8.9%	2.0%	4.3%	75.6%	9.9%
Nancy	346	38.2%	4.0%	7.8%	31.5%	18.8%
Helen	933	42.7%	12.8%	17.3%	14.9%	12.4%
Henry	897	31.5%	9.7%	15.2%	42.1%	1.6%
Edward	177	24.3%	10.2%	27.1%	36.2%	5.1%
Linda	117	23.9%	17.9%	18.8%	32.5%	6.8%
Mean		23.7%	9.4%	13.4%	42.9%	11.0%
OAT	76	18.4%	19.7%	18.4%	26.3%	21.1%
Indicators	51	15.7%	19.6%	11.8%	31.4%	21.6%

Table D8

2005-06 Assessment Item Data: Below Grade Level Items

		Standards					
Teachers	N	Number	Measure	Geometry	Algebra	Data	Totals
Sam	91	4.4%	3.3%	6.6%	2.2%	4.4%	20.9%
Wanda	223	11.7%	7.6%	11.2%	1.3%	4.9%	36.8%
Frank	146	5.5%	4.1%	3.4%	2.7%	1.4%	17.1%
Evelyn	393	5.3%	0.3%	3.1%	3.6%	1.0%	13.2%
Nancy	346	24.0%	2.3%	2.3%	3.8%	6.1%	38.4%
Helen	933	36.0%	8.3%	12.0%	1.9%	4.8%	63.0%
Henry	897	21.9%	5.7%	9.4%	2.2%	0.8%	39.9%
Edward	177	13.0%	5.1%	14.1%	0.0%	3.4%	35.6%
Linda	117	10.3%	9.4%	7.7%	0.0%	0.0%	27.4%
Mean		14.7%	5.1%	7.8%	2.0%	3.0%	32.5%

Table D9

2005-06 Assessment Item Data: Items Above the Eighth Grade

		Standards					
Teachers	N	Number	Measure	Geometry	Algebra	Data	Totals
Sam	91	0.0%	0.0%	0.0%	18.7%	1.1%	19.8%
Wanda	223	0.0%	0.0%	0.0%	8.5%	2.2%	10.8%
Frank	146	0.0%	0.7%	0.0%	3.4%	2.1%	6.2%
Evelyn	393	0.5%	0.0%	0.0%	42.2%	1.5%	44.3%
Nancy	346	0.0%	0.0%	0.0%	2.3%	1.4%	3.8%
Helen	933	0.4%	0.0%	0.6%	0.6%	1.7%	3.4%
Henry	897	2.1%	0.0%	3.1%	22.4%	0.0%	27.6%
Edward	177	0.6%	0.0%	0.0%	11.9%	0.0%	12.4%
Linda	117	0.0%	0.0%	0.0%	0.9%	0.0%	0.9%
Mean		0.4%	0.1%	0.4%	12.3%	1.1%	14.3%

Table D10

2005-06 Assessment Item Data: Depth of Knowledge

		Depth of Knowledge		
Teachers	N	Level 1	Level 2	Level 3
Sam	91	58.2%	38.5%	3.3%
Wanda	223	83.9%	16.1%	0.0%
Frank	146	82.9%	13.7%	3.4%
Evelyn	393	93.1%	6.6%	0.3%
Nancy	346	91.0%	8.7%	0.3%
Helen	933	97.7%	2.1%	0.1%
Henry	897	98.8%	1.2%	0.0%
Edward	177	92.1%	7.3%	0.6%
Linda	117	79.5%	19.7%	0.9%
Mean		86.4%	12.7%	1.0%
OAT	76	61.8%	35.5%	2.6%